Using online search data to forecast new product sales

Gauri Kulkarni a, P.K. Kannan a,⁎, Wendy Moe b

a Department of Marketing, Sellinger School of Business and Management, Loyola University Maryland, Baltimore, MD 21210, USA
b Department of Marketing, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, USA

A R T I C L E   I N F O

Article history:
Received 19 April 2011
Received in revised form 1 August 2011
Accepted 13 October 2011
Available online 25 October 2011

Keywords:
Online search
Entertainment marketing
New product sales forecasting

A B S T R A C T

Search engines are rapidly emerging to be the “go-to” sites for consumers to learn more about a product, concept or a term of interest, irrespective of the initial channel in which the interest originated — text, radio, TV, multimedia channels, word of mouth, etc. In this paper we argue that data on the search terms used by consumers can provide valuable measures and indicators of consumer interest in a product, concept or a term. Such data can be particularly valuable to managers in gauging potential product interest in a new product launch context or consumption interest in the post-release context. Based on this premise, we develop a model of pre-launch search activity and link the pre-launch search behavior and product characteristics to early sales of the product, thus providing a useful forecasting tool. Applying the model in the context of motion pictures, we find that search term usage follows rather predictable patterns in the pre-launch and post-launch periods and the model provides significant power in forecasting release week sales as a function of pre-release search activity. With advertising data included in the model, we find that the pre-release search data offers additional explanatory and forecasting power, thus highlighting the ability of the search data to capture other factors, such as possibly word-of-mouth, in impacting early sales. We offer specific insights into how managers can use search volume data and the model to plan their new product release.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

With the advent of the Internet, consumers are able to search for information about virtually anything with the click of a button, often in the comfort of their own homes. The Internet has dramatically lowered the cost of information search [1], and there has been much work done on this [18,23]. According to recent research, search engine use is a major activity of people who use the Internet, and data on terms that are searched for can be easily obtained. This data is often collected for search engine advertising purposes. It can also have several other useful applications that have not received much attention in the marketing literature. Of particular interest is its use as a measure of word-of-mouth, buzz, effect of advertising, etc.,— or overall consumer interest.

Both Internet use and search engine use are becoming increasingly common in the United States. Internet penetration in the United States has hit an all-time high with 73% of adults reporting Internet use and 65% of users reporting daily use [14]. Of the Internet users, 91% report using a search engine to find information. This activity is second out of all Internet activities, with using the Internet to send or read e-mail as the most common online activity [14]. Thus, it is clear that both Internet use and search engine use have become a major part of adult life in the United States. The same report finds that the most common search terms are related to pop culture, news events, trends, and seasonal topics [10]. The entertainment or recreation category was sixth in terms of number of online queries in 2002 [10]. Thus, it is reasonable to believe that many consumers search for new product information online using a search engine.

The search literature has differentiated between consumer search for product information when (1) they seek knowledge on specific attributes of a product and (2) they seek knowledge on how a particular product compares relative to others [16,22]. Prior research has found that consumers will still benefit from search if they have knowledge on the offerings of a product but are uncertain about how that product stands relative to others when making a choice. In the case of new products, it is very likely that consumers are uncertain about choice of a product relative to others, since they have no prior experience with the product. Thus, consumer information search for a new product is likely to take place.

Information search for new products can take place in several forms including consultation of product information sources (e.g. manufacturer print/web sources), third-party sources (e.g. consumer reports), word-of-mouth, or use of a search engine to look for information from a variety of sources on the Internet. Search engine use requires the entry of a search term for the topic at hand — in our research context, the name of the product of interest. Thus, the search we focus on is search term volume, or the aggregate number of times particular terms are submitted to online search engines. Given the characteristics of search terms and search engine use that are mentioned earlier, we argue that this measure can capture interest in

⁎ Corresponding author. Tel.: +1 301 405 2188; fax: +1 301 405 0146.
E-mail addresses: gm.kulkarni@loyola.edu (G. Kulkarni), pkannan@rhsmith.umd.edu (P.K. Kannan), wmoe@rhsmith.umd.edu (W. Moe),

0167-9236/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2011.10.017
current trends in pop culture and new products. In other words, the terms that consumers choose to submit to search engines indicate their interest or concern for specific topics or products [9], and the overall volume of the terms can indicate the general (aggregate) level of interest. Therefore, we propose the use of this measure of consumer interest in new products to predict new product sales.

The research objective of this paper is to introduce a new measure of consumer interest and evaluate the predictive power of this measure in new product sales forecasting. We aim to answer two questions: (1) Is search term volume a good measure of consumer interest? and (2) Does this measure offer good forecasting power? To the best of our knowledge, we are the first to consider search term volume as a marketing metric. While several recent research studies in marketing have examined various sources of online word-of-mouth, including online conversations, reviews, opinion platforms, blogs, etc. [20], we propose a measure that offers several advantages.

One major advantage of search term data is that one is able to search for a product prior to its launch. Thus, the measure can be obtained during a product's pre-launch phase, allowing for pre-launch forecasting, a task that managers have long struggled with. Search engine use is also a more prevalent online activity than participation in newsgroup conversations, writing online reviews, or blogging. Therefore, search engine data is more likely to be representative of the general online population. Search term volume also does not require analysis of content or great amounts of data cleaning or coding, making it an attractive measure for managers to work with. It can also be collected or obtained with ease and little cost. Changes in search term volume over time also allow for examination of trends or patterns in consumer interest. Thus, we argue that search term volume offers several advantages to many online measures that have recently been used in new product sales forecasting.

Our research contributes to the existing literature on online information search and new product sales forecasting. While we choose to illustrate our proposed framework in the motion picture industry, our approach can be generalized to other product categories such as books or music. For example, search volume may be used to predict release-week sales for new music albums or books. The pattern and level of search may also be useful in predicting sales in later weeks, since these products tend to have longer product life cycles than most motion pictures. We find that both search volume and search volume pattern over time are important in predicting box-office sales. In doing so, we have developed a forecasting approach that will aid managerial decision making for new products.

2. Conceptual development

We illustrate our framework by forecasting motion picture box office revenues using online search engine activity. We develop a forecasting approach that models pre-launch and post-launch search, and ultimately forecasts sales. We focus on opening weekend sales only, since these are most critical and also most difficult to predict [6]. We posit that pre-launch search is largely driven by consumer interest in a product, while post-launch search is driven by interest in both the product, as well as consumption of the product. Product interest refers to interest in specific attributes of the product, while consumption interest refers to interest in consuming the product. Thus a consumer may search for information regarding actors/actresses, trailers, plot summary, etc. during the pre-launch phase of a motion picture. A consumer may search for this information during the post-launch as well, but he may also search for information about theater locations, show times, consumer reviews, etc. during the post-launch phase. We incorporate product characteristics in our model, such as genre and MPAA rating, as these are likely to have effects on both search and sales. We also control for competition.

In our base framework, we do not account for any drivers of search. Therefore, a possible alternative explanation for the predictive power of search term volume is that it is capturing the response to advertising, and since advertising expenditure data is also available during the pre-launch period, search term volume does not offer a significant gain. In 2006, theater admissions, ticket prices, number of movies released, box-office revenues, and production costs were up from 2005 [17]. The average production cost per film for MPAA member companies was $65.8 million, while the average marketing cost was $34.5 million [17]. These figures suggest that motion pictures remain a growing industry in the entertainment category, and marketing expenditures play an important role in the success of motion pictures. Therefore, we extend our analysis and examine the role of advertising in our framework. Modeling advertising expenditures allows us to determine whether search term volume is capturing the effect of advertising or consumer interest generated from other sources, such as word-of-mouth, above and beyond advertising. It also allows us to compare forecasting performances for modeling approaches with and without advertising.

In our study, online search term volume refers to the aggregate number of times a movie title is searched for using an online search engine. We posit that online search term volume offers a main advantage to measures such as online reviews [3,11,13]. Unlike posting a review, a consumer need not have viewed the movie to search for it online. Thus, this activity can be captured several weeks before the movie is released in theaters — the pre-release period. Additionally, since virtually anybody can search for a movie title online, effects of the distribution or availability of the movie do not come into play as much. In other words, anybody can search for a movie regardless of when, where, or how often it is showing. Although the search volume does not capture any content or valence as a review does [4], since Liu [13] found that volume offers significantly more explanatory power than valence, we posit that the advantage of having more pre-release data outweighs this. We argue that searches indicate consumer interest in a product, and thus we aim to use this measure of interest to forecast sales. We discuss the components of our framework in detail in the next sections.

2.1. Internet and search engines

Many would agree that the Internet is at the core of the recent technological revolution. The age of information is associated with the ease of availability of information, largely due to the Internet. Consumers are able to search for information about virtually anything using the Internet. The most basic tool that enables this is an online search engine [14]. Since an online search requires an action on the part of the consumer, i.e. entering a search term, there must be a trigger or driver of any given search and the desire for more information. This driver could be WOM, advertising, promotion, news coverage, or any combination of these. Thus, one could suggest that online searches are a measure of interest or “buzz” for a topic or product.

An interesting next step would be to investigate the relationship between these searches and actual consumption of products. In other words, can the number of searches for a given product suggest the level of interest in that product and therefore help predict sales for that product? Given that searches tend to relate to pop culture and trends [10], this type of relationship is likely to be strongest for products that are new or “trendy” and have heavy pre-launch marketing campaigns. These would include high-technology products, music, and movies. These types of products tend to generate high levels of WOM and often have heavily publicized launch dates. Thus, we aim to use searches for a new product to predict the sales of that product.

Search term data is easily collected and obtained. It also does not require as much data cleaning or coding as analyses of online messages or conversations, thus it can be used on a larger scale. Additionally, since search engine use is a very prevalent online activity as
compared to blogging, participation in newsgroups, etc., the data is less likely to suffer from selection bias and is more representative of the general online population. Search term volume is able to measure consumer interest very early in the consumer decision process (before the consumer has made a purchase decision), as well as in early stages (pre-launch) of the product life-cycle. We explore this stage and its implications for forecasting in a digital context.

2.2. Motion picture context

We choose the motion picture industry to explore this idea because movies have relatively short life-cycles and reliable data is easily available. There are also several sources of movie-related information available online [7]. Additionally, motion pictures involve high levels of pre-launch marketing activity that generate consumer interest. Therefore, consumers are able and likely to search for motion picture information prior to release.

Eliashberg, Elberse, and Leenders [7] give several examples of movie-related information sources that are available online. These include chat rooms, Web logs, portals, recommendation sites, customer and critic review sites, official movie sites, and databases. Additionally, consumers may search movie titles to find show times or locations of theaters screening a particular title. Thus, we argue that consumers who are interested in viewing a movie may search for it online using a search engine.

The first week of release is often the most crucial for motion picture revenues, and is also the most difficult to forecast. Therefore, pre-launch forecasting will provide several useful implications to managers for scheduling, resource allocation, forecasting, etc. [2,8]. Given the characteristics of online search terms that are mentioned earlier and these characteristics of the motion picture industry, we think it is a good area to begin investigating the potential of search term activity as a forecasting measure. While we use motion pictures to illustrate our forecasting framework, we emphasize that our approach is generalizable to other new product launches, particularly those that involve high levels of pre-launch marketing activity and heavily advertised launch dates. Specifically, search term volume and pattern over time can be used as a similar measure of consumer interest for products such as books or music, and our framework can be applied to forecast release-week sales and extended to predict sales in later weeks.

2.3. Pre- and post-launch

We consider search for a product in two phases—pre-launch and post-launch. We posit that pre-launch search is largely driven by consumer interest in the product. Consumers may search for general information about the product, features about the product, press releases, etc. Pre-launch information search is likely to be an indication of product interest. This could be a response to WOM, advertising, promotion, or other media coverage related to the product’s upcoming launch. Post-launch search would also include product interest, however it will also be affected by interest in consumption of the product. Thus, after a product is launched, consumers may search for a product with interest in finding information about availability, release times, etc. This is in addition to the increased WOM or buzz that is likely to take place after a product is launched [13], since other consumers have now purchased the product and are able to talk about it. This increase in WOM is also likely to result in increased online search.

2.4. Conceptual framework

We illustrate our conceptual framework in Fig. 1 and discuss it in detail in the next section.

2.4.1. Pre- and post-release search (product and consumption interest)

We begin with pre-release search. We argue that this indicates consumer interest in the product. The sources of this interest can include exposure to WOM, advertising, promotion or press coverage, although we do not explicitly measure any of these in the current framework. Given that the product it not yet available for consumption or purchase, a search for it suggests some level of interest in it. For example, a consumer that searches for a movie before it is released is likely to be interested in obtaining information specific to the movie, such as details about actors/actresses, plot or story line, trailer, etc., although the movie is not yet available for purchase or consumption.

Once the product is launched or released, online search will indicate interest in both the product itself, as well as in consumption of the product. In our illustration, consumption interest could drive consumers to search for a movie to find information about show times, theater locations, consumer reviews, etc. This information is not likely to be available during the pre-release period.

We distinguish between pre- and post-release search in our conceptual development, as the motivations to search and therefore the types of consumer interest that are being captured may differ in these two stages. However, the distinction is not critical to our model, and therefore pre- and post-release search are not treated differently in our modeling framework.

2.4.2. Purchase (box-office sales)

We further link search to box-office sales. Since we are using search activity as a measure of interest, this interest should translate into purchase of the product. We argue that it is reasonable to believe that the levels of consumption interest and product interest will have a strong relationship with ultimate purchase.

2.4.3. Product characteristics

Characteristics inherent to the product are the main attributes that potential consumers are interested in. We argue that product characteristics can have an impact on both search and sales. In the case of motion pictures, product characteristics include attributes such as genre, MPAA rating, and production budget. Consumers are likely to have preferences for particular genres or MPAA rating levels of movies. These types of attributes are likely to affect consumers’ interest in a given movie, and will thus impact their propensity to search for more information about a given movie, in both the pre- or post-launch phases. They will also play a role in a consumer’s decision to view a given movie once it is released [21].

2.4.4. Competition

It is also important to control for competition or the number of alternatives available for purchase [13]. Competition will also affect both search and sales. Consumers are not likely to search for information about every movie in the market. Thus, the more movies there
are in the market, the less likely a consumer will search for any given alternative. The same rationale extends to sales — the more competition there is, the fewer the sales for any given movie.

Our framework aims to link product interest, consumption interest, and product sales while controlling for the effects of competition and product characteristics. We propose online searches as a measure of product and consumption interests and examine their relationship with sales.

3. Model development

Our objective in this paper is to model pre- and post-release search behavior and relate observed patterns of search to opening week box office sales. Search can be characterized by (1) the total volume of search activity observed and (2) the pattern of search activity over time. We model the total volume of search for movie \( i \) as follows:

\[
\ln(\text{Search Volume}_{it}) = \alpha_i + fX_i + \epsilon_i
\]

(1)

where \( fX_i \) represent movie-specific covariate effects, \( \epsilon_i \) captures heterogeneity across movies, and \( \epsilon_i \sim \text{normal}(0, \sigma) \). The total search volume is over the period of pre-release duration under study.

We model the week-to-week pattern of search activity using a Weibull hazard process. As seen in Fig. 2 below, the search pattern for movies can be very different, varying in their slopes and shapes.

We choose the Weibull for its flexibility in capturing various shape patterns, such as the ones we observe in the search data. The hazard model specification is ideal for this context as it captures the probability that a search will take place in time \( t \) given that the search has not taken place by \( t-1 \). Aggregating provides a model of total searches in each week. We specify the hazard function as follows:

\[
\text{Hazard} \ : h_i(t) = \frac{f(t)}{1-F(t)} = \lambda_i c t^{\gamma_i} \exp(\gamma Z_i)
\]

(2)

where \( \lambda \) and \( c \) are movie-specific parameters to be estimated, \( t \) indexes time and \( \gamma Z_i \) represents covariate effects which we will specify later in Section 4.1. The pattern is captured through \( \lambda \) and \( c \), where \( \lambda \) is the slope parameter and \( c \) is the shape parameter. In general, our search data shows a non-linear growth trend as the week of launch approaches, and the Weibull performs well in modeling this pattern over time.

Finally, we model box office sales by focusing on release week sales. Previous research has shown that sales for hedonic products (e.g., movies and music) follow rather predictable patterns [15]. The data shows a non-linear growth trend as the week of launch approaches, and the Weibull performs well in modeling this pattern over time.

\[
\ln(\text{Sales}_{it}) = \alpha_i + bX_i + u_i \quad \text{where} \quad u_i \sim \text{Normal}(0, s).
\]

(3)

Thus far, we have specified the search volume, search pattern, and sales volume models separately. However, these constructs are related. That is, there is inherent endogeneity in these measures. For example, consumers who have a lot of interest in the details of a movie may contribute to a large search volume as well as large sales for the movie. Additionally, a spiked pattern in search activity just before movie release may have different impact on first-week sales figure as against a slow build of pre-release search activity. Thus, the search activity pattern that we model using a hazard model may also be correlated with sales once the movie is released. To account for these interference effects among the three models, we let the parameters of the model to be jointly correlated and determined by the hyper-parameters we define below:

\[
\begin{bmatrix}
\alpha_i \\
\beta_i \\
\ln(\lambda_i) \\
\ln(c_i)
\end{bmatrix}
\sim \text{MVN}(\omega, \Sigma)
\]

(4)

where \( \omega \) and \( \Sigma \) are the mean vector and covariance matrix (the hyperparameters). This model allows us to flexibly relate the search volume and slope and shape parameters of the pre-release search patterns to the baseline sales of each movie \( i \), accounting for the endogeneity across these models. This treatment also allows us to theoretically improve the forecasting accuracy of our model. The covariates used in the model are described in the following section.

4. Data and estimation

We create our dataset by merging three types of data — search data, movie-related variables, and advertising expenditures. We obtain the search data from a search term research service that collects and compiles search term data. The service maintains a database of search terms collected from all of the major search engines, including Google, Yahoo!, and MSN. Search term volume is available at the weekly level and refers to the number of times the term is searched for within a given week. For example, for the term “ipod,” one could obtain the number of searches on that term for a given week. The data comes from a database based on user panel data and is free from skew caused by automated agents. It contains data on over 4.3 billion searches. We obtain the data at the aggregate level and do not have demographic information on the user panel. Specifically, we obtain the weekly search volume for each movie title in our analyses during both the pre-launch and post-launch phases. In other words, we have the number of times each movie title is searched for on a weekly basis. Our analysis focuses on 9 weeks of search data — 8 weeks of pre-launch and the week of launch (1 week of post-launch).

Our dataset includes movies released from July to November of 2006 in the United States. There were 599 new feature films released in the US in 2006, and 63 grossed more than $50 million in box office revenues [17]. We limit our set of titles to those that were in the top fifty in box office revenues during their opening week. After removing titles for which we were not able to obtain all variables or had very sparse search data, we conduct our analyses on 61 movie titles. We obtain motion picture data from two popular movie sites, The Numbers (http://www.the-numbers.com) and Yahoo Movies (http://movies.yahoo.com/).

---

**Fig. 2. Differences in Search and Sales Patterns for Two Movies.**
4.1. Covariate specification

We include covariates for competition, production budget, genre, and MPAA rating in our vector of X variables and measures of advertising expenditures in Z. We collect box office release date, weekend box office revenues, MPAA rating, and genre for each movie title in our analyses. We use the box-office release date to determine the competition level. We define Competition as the number of other movies in our dataset that are released in the same week as a given movie and include it as a covariate. We collect production budget data from The Internet Movie Database Pro (http://www.imdb.com/). We include this as covariate Budget.

We have three categories of genre. Comedy refers to comedy or romantic comedy movies. ActAdvHor refers to movies that are action, adventure, or horror. Drama represents drama. We include Comedy and Drama as indicator covariates in our models (ActAdvHor is excluded). We also have three categories of MPAA ratings, PG, PG13, and R. We include PG and PG13 as indicator covariates as well, excluding R. Our dataset does not contain any movies rated G or NC17. We obtain both genre and MPAA rating data from the above-mentioned web-sites.

Our dataset also includes advertising expenditures for each of the movies for both the pre- and post-launch periods on a weekly basis. The expenditures reflect the week that the advertising occurred and not the week that the payment was made. The data includes advertising expenditures across television, radio, magazines, newspapers, Internet, and outdoors. We obtain the advertising data from a commercial media and advertising database. Similar to our approach in the search volume framework, we consider advertising expenditures up to eight weeks prior to the motion picture’s release. We include this as a time-varying hazard covariate in Z in our model.

In our analyses, we focus only on the opening weekend box office sales. We do so because box office sales tend to follow fairly predictable patterns in subsequent weeks, and these can often be determined from the first week’s sales.

4.2. Estimation

We employ hierarchical Bayesian methods to estimate the proposed model[18] and use a freely available software program, WinBUGS. WinBUGS employs MCMC techniques that simulate multiple iterations of parameter estimates until the sampler converges to a solution (see [15]). We obtain 10,000 iterations, discarding the first 5000 iterations as burn-in and using the last 5000 to provide parameter estimates. The results provided in the next section represent parameter medians and confidence intervals in these latter 5000 iterations.

5. Results and discussion

We compare fit and performance for four models: an independent model, a correlated model without movie-specific covariates in the search volume component (but included in the sales component), a correlated model with movie-specific covariates in both the sales and search volume components, and an advertising effects model.

First, we estimate an independent model that does not correlate the model parameters for search and sales to serve as a baseline (Independent). This specification assumes that search and sales are independent of one another. Having this model to serve as a baseline will highlight the importance of allowing for the search and sales parameters to correlate.

Next, to account for the interdependence between search and sales, we correlate their model parameters and estimate two versions of the model. The first does not include movie-specific covariates in the search volume component (Correlated No Covar), while the second does incorporate movie-specific covariates in the search volume component (Correlated With Covar). Both of these models include movie-specific covariates in the sales component. The purpose of these two models is to examine the importance of movie-specific effects on search volume or search propensity for specific types of movies. In other words, these two specifications allow us to determine whether movie-specific covariates have an effect on search volume.

Lastly, we incorporate the effect of advertising to account for one driver of search (Ad Effect). We present our results in Tables 1 and 2. Table 1 gives an overview of model fits.

5.1. Model fits

Table 1 presents statistics on model fit. The deviance information criterion (DIC) is used to compare hierarchical Bayesian models. Smaller DIC values indicate better fit. The average percentage error (APE) measures the difference between actual and predicted ln(sales). The results indicate the importance of correlating the search and sales parameters, as all of the correlated models perform better than the Independent. We can see that movie-specific covariates (Correlated With Covar) improve the fit of search but not of sales. This implies that it is the top line search that is important and not the baseline search propensity. We can also see that Ad Effect provides dramatically better fit of search pattern. The improvement in the sales model is significant but less dramatic. These results highlight the importance of incorporating advertising expenditures into the modeling framework.

5.2. Results

Table 2 presents the estimation results for the three (correlated) model specifications, without movie-specific covariates in search, with movie-specific covariates but without advertising effects in search, and with both movie-specific covariates and advertising effects in search.

First, we can observe some general patterns. Dramas tend to generate lower sales at the box office, while moves rated PG-13 tend to generate more sales. We see a positive and significant advertising effect on search pattern.

After movie-specific covariates are included in the search model (middle columns), we can see that the effect of production budget on ln(sales) becomes significant. The positive effect implies that movies with higher production budgets result in higher sales. Movies with larger production budgets also generate higher search volume.

Table 1
Model fits.

<table>
<thead>
<tr>
<th></th>
<th>Independent</th>
<th>Correlated no covar.</th>
<th>Correlated with covar.</th>
<th>Ad effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIC</td>
<td>2087.39</td>
<td>2054.67</td>
<td>2064.24</td>
<td>2056.83</td>
</tr>
<tr>
<td>ln(sales)</td>
<td>1115.07</td>
<td>1091.59</td>
<td>1089.50</td>
<td>1092.54</td>
</tr>
<tr>
<td>ln(searchVolume)</td>
<td>61,485.10</td>
<td>61,481.20</td>
<td>61,479.50</td>
<td>51,356.70</td>
</tr>
<tr>
<td>Search</td>
<td>64,687.55</td>
<td>64,627.50</td>
<td>64,633.30</td>
<td>54,506.10</td>
</tr>
<tr>
<td>Total</td>
<td>608</td>
<td>608</td>
<td>608</td>
<td>608</td>
</tr>
<tr>
<td>APE (in-sample)</td>
<td>3.99</td>
<td>2.05</td>
<td>2.79</td>
<td>2.29</td>
</tr>
</tbody>
</table>
which is not surprising. The “big” movies tend to result in higher levels of consumer interest. We can see that comedies result in lower search volume, as do movies with higher levels of competition, suggesting lower levels of interest.

We can further see that after the advertising effect is included (right columns), the movie-specific covariates in the search model (comedy, competition, production budget) are insignificant. The production budget effect on ln(sales) is also insignificant. This is captured by the advertising effect which then correlates with baseline sales. The PG effect on ln(sales) becomes significant and is positive.

5.3. Forecasting

Lastly, we forecast sales using our modeling framework. We divide our sample of 61 movies into calibration (40 movies) and validation (21 movies) samples. Thus, we estimate our models on 40 movies and predict sales of the remaining 21 movies. We forecast using all three (correlated) model specifications (no covariates in search, no advertising effects, and with advertising effects) for two calibration periods — 8 weeks of pre-launch search data (forecast provided 1 week prior to launch) and 4 weeks of pre-launch search data (forecast provided 1 month prior to launch). We compare our forecasting performance to a baseline forecast which includes no pre-launch search (only movie characteristics). Our results are in Table 3.

From the forecasting results, we can see that forecasting performance is greatly improved with the inclusion of pre-launch search. Further, we can see that forecasting performance is better with the four-week calibration period than the eight-week calibration (which includes all 8-week search data, one week prior to release). This suggests that consumer interest measured one month prior to launch is a better indicator of sales than consumer interest measured just 1 week prior to launch. Across both calibration periods, we can see that forecasting performance is best with the specification that does not include movie-specific covariates in the search model.

5.4. Discussion

The primary focus of this study is to investigate the predictive power of online search data in forecasting new product sales, movies, in our specific case. The forecasting results of our model indicate that search data does indeed offer predictive power in forecasting sales. The APE is under 10%, suggesting that search data may offer a useful measure to managers interested in forecasting sales of a new product, particularly in the pre-launch period. Thus, our approach can be applied to a product's launch, offering managers a powerful forecasting system. As discussed earlier, the forecasting of new product sales early in their life-cycle is a problem that managers have long struggled with. Our framework offers one possible approach to address this issue, and allows managers to build useful decision support tools.

This work contributes to the extant stream of research on measures that can be used to forecast product sales. Recent research has considered measures such as online reviews, “tweets,” advance purchase orders, and early sales data, which have also been found to offer predictive power [4,13,15,20,21]. Our work extends the recent focus on electronic measures by offering a new measure and applying it in the context of motion pictures. While online reviews have been successfully used to forecast box-office revenues [3,4,13], consumer reviews are often not posted until after a movie is released. Thus, forecasting during the pre-launch period is not possible. As mentioned earlier, the opening weekend is the most critical and the most difficult to predict, as revenues in later weeks often depend on the first week's performance [6]. Therefore, our proposed measure provides a valuable indicator of consumer interest which is available during the pre-launch period.

In this research, we set out to answer two questions: (1) Is search term volume a good measure of consumer interest? and (2) Does this measure offer good forecasting power? Our results indicate that yes, search term volume is a good indicator of consumer interest and yes, this measure does offer forecasting power. Thus, our main contribution in this study is two-fold — we offer a new measure of pre-
launch or pre-release consumer interest and apply it to a forecasting framework. The overall results and managerial implications of our study, as well as concluding remarks and areas for future research, are discussed in the next section.

6. Conclusion

Our main objective is to illustrate the effectiveness of online search volume as a new product sales forecasting measure. We are particularly interested in the pre-launch aspect of forecasting. We illustrate this in the context of motion picture revenues. We distinguish between online search that takes place before launch and search that takes place after launch (post-launch). We use pre-launch search volume as a measure of product interest and post-launch search as a measure of both consumption interest and product interest. Post-release search can be driven by product interest, as well as interest driven by other characteristics such as product availability.

We develop a modeling framework that links pre- and post-launch search, and box-office revenues. We also incorporate product characteristics, including competition. We extend our framework and model the effect of advertising. Doing so allows us to account for at least one driver of online search and compare the forecasting performances of our modeling approaches. We find that online search is a significant predictor of opening-weekend box-office revenues. We also find that our framework performs well as a forecasting tool.

6.1. Managerial implications

Our first contribution lies in the proposal of a new measure of consumer interest that is available prior to a new product’s launch. WOM is one indication of consumer interest, and measurement of WOM is an issue that has been raised in the literature on WOM. Therefore, we introduce an easily-obtained, cost-effective measure for consumer interest in a new product. Data on search activity is easy to collect and clean. It does not require much coding or analysis of content. Search engine activity is also a very prevalent online activity as compared to blogging or writing consumer product reviews. The data is available early in the product’s life cycle, and can capture consumer interest in a new product early in the consumer decision process. In other words, search data is available before a new product is launched, often several months prior to launch. Therefore, this measure can provide useful measures of consumer interest in a new product well in advance of the product’s launch. This gives managers the opportunity to adjust their marketing strategy as necessary, depending on the levels of consumer interest in the new product before the product is even available for consumption.

Our second contribution is in the development of a model that uses this measure to forecast new product sales. We first develop a model that forecasts sales using product characteristics and search term data. We then extend our modeling framework to include one possible driver of search activity — advertising. We focus on advertising because we are illustrating our framework in the context of motion pictures, where advertising expenditures can play a large role in the success of a movie. We find that search data offers predictive power in forecasting opening-weekend box-office revenues, and our modeling framework and forecasting procedure perform quite well. These results have useful implications for managers of new products. Managers can monitor the search pattern and volume for terms related to their products during the pre-launch period to gauge interest in their new products. They can use this information to alter their marketing strategy, if necessary.

Our study offers several opportunities for further research in this area. We discuss limitations of our research and areas for future work in the next section.

6.2. Limitations and future research

One problem that is faced when using search data is the lack of very clean data. For example, some movie titles that are also related to other products (e.g. water, firewall, cars) may be searched for using search terms that involve more than just the title. Thus, the data for these terms may be contaminated. Also, sometimes a motion picture title is also a book title. If a movie title is very long, perhaps online searchers only enter the first few words or the main words as search terms. Consumers may also search using the names of actors or actresses with roles in the motion picture. Thus, in some cases, it may be difficult to tell exactly what a consumer would enter as a search term when searching for information on a motion picture. Some experimental work in this area would help to better understand consumer choice of a search term.

Although this work represents an example of the usefulness of online search volume as a predictor of motion picture success, there exist many opportunities for future research. A similar problem could be examined for DVD release dates. Also, with the historical data on search volume, other times of high motion picture interest could be determined. For example, DVDs are often given as gifts during the holiday season. Thus, there is likely to be a surge in searches for particular titles during this time. As mentioned earlier, the most recent work in this area has focused on online reviews and ratings posted by consumers. An obvious extension would be to combine reviews/ratings and search volume into a consumer interest measure in a forecasting model.

Also, the timing problem could be examined. There has been research done on the optimal timing of a motion picture release on DVD based on the success of the motion picture in theaters [12]. Perhaps WOM data captured by search volume could help optimize this solution as well. Elberse and Eliashberg [5] look at the sequential release of motion pictures in international markets. They find that “the longer is the time lag between releases, the weaker is the relationship between domestic and foreign performance.” Though they focus on screen allocations, the authors suggest that this is “consistent with the idea that the ‘buzz’ for a movie is perishable.” If this is the case with international markets, it is likely to be the case for the box-office — DVD performance relationship. Again, perhaps the strength of this “buzz” could be captured by search volume to help optimize this solution as also.

Our approach could also be extended to product categories beyond motion pictures. Search data on terms related to other products are just as easily available. Thus, our framework could be applied to products such as books, music, or technology products. Our framework is flexible in the sense that the probability distributions that are used can be adapted to fit the data better. Other distributions may be used for products with longer life-cycles or different patterns of search, in general.

Lastly, there are some demographic biases when looking only at Internet data. For example, it has been reported that males tend to use the Internet more than females. Also, Internet use tends to increase with household income and education level. People of Hispanic ethnicity are least likely to use the Internet [14]. Additionally, younger Internet users are more likely to use search engines and use them often [10]. We do not address these issues in our study.

While there are limitations and opportunities for further study in our area of research, we take the first step of measuring consumer interest in a new product using online search term data and use this data to successfully predict new product sales.

References


Gauri Kulkarni is Assistant Professor of Marketing at the Sullinger School of Business and Management at Loyola University Maryland in Baltimore, Maryland. The present research is based on her doctoral dissertation in the Robert H. Smith School of Business at the University of Maryland. Professor Kulkarni’s current stream of research focuses on consumer information search – both online and offline. She has developed new product sales forecasting models based on online search data. She has also considered the impact of search channel and information sources on consumer choice.

P. K. Kannan is Ralph J. Tyser Professor of Marketing Science, Robert H. Smith School of Business, University of Maryland, College Park, Maryland, and he is the Chair of the Department of Marketing. His current research stream focuses on new product/service development, design and pricing of digital products and product lines, marketing and product development on the Internet, e-service, and customer relationship management (CRM) and customer loyalty. He has received several grants from National Science Foundation (NSF), Mellon Foundation, SAIC, and PricewaterhouseCoopers for his work in this area. His research has also won the John Little Best Paper Award (2008) and the INFORMS Society for Marketing Science Practice Prize Award (2007). Professor Kannan’s research has also been selected as a finalist for the Paul Green Award (2008).

Wendy Moe is Associate Professor of Marketing, Robert H. Smith School of Business, University of Maryland, College Park, Maryland. She is an expert in the area of online behavior and early sales forecasting. Her research has focused on developing statistical methods and models for Internet clickstream data, online social media/user-generated content and entertainment sales (e.g., sales of music, event tickets, etc.). Professor Moe has been recognized by the American Marketing Association as the Emergent Female Scholar and Mentor in the field of marketing with the 2010 Erin Anderson Award. She has also worked with several firms to develop and implement state-of-the-art statistical models in the context of web analytics and product forecasting.