

Blogs as Predictors of Movie Success

Eldar Sadikov and Aditya Parameswaran and Petros Venetis

Computer Science Dept., Stanford University, Stanford, CA 94305

esadikov,adityagp,venetis@stanford.edu

Abstract

Analysis of a comprehensive set of features extracted from blogs for prediction of movie sales is presented. We use correlation, clustering and time-series analysis to study which features are best predictors.

Introduction

In this work, we attempt to assess if blog data is useful for prediction of movie sales and user/critics ratings. Here are our main contributions:

- We evaluate a comprehensive list of features that deal with movie references in blogs (a total of 120 features) using the full `spinn3r.com` blog data set for 12 months.
- We find that aggregate counts of movie references in blogs are highly predictive of movie sales but not predictive of user and critics ratings.
- We identify the most useful features for making movie sales predictions using correlation and KL divergence as metrics and use clustering to find similarity between the features.
- We show, using time series analysis as in (Gruhl, D. et. al. 2005), that blog references generally precede movie sales by a week and thus weekly sales can be predicted from blog references in the preceding weeks.
- We confirm low correlation between blog references and first week movie sales reported by (Mishne, G. et. al. 2006) but we find that (a) budget is a better predictor for the first week; (b) subsequent weeks are much more predictive from blogs (with up to 0.86 correlation).

Data and Features

The data set we used for this paper is the `spinn3r.com` blog data set from Nov. 2007 until Nov. 2008. This data set includes practically all the blog posts published on the web in this period (approximately 1.5 TB of compressed XML).

Since our focus is on prediction of movie success, we identified the following relevant output variables:

- Average critics ratings
- Average user (viewers) ratings
- 2008 gross sales

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- Weekly box office sales, weeks 1–5

The first two of these variables were collected from `rottentomatoes.com` and the last two were collected from `the-numbers.com`.

We then collected a list of the top 300 Box Office movies of 2008 from `the-numbers.com` and manually filtered the list for the titles that were similar to the common English phrases. For example, searching for ‘Wanted’ or ‘21’ triggered many false positives that might not refer to the movies. Since such titles could have skewed our results in a non-uniform way, we eliminated such titles. For the remaining 197 movies, we constructed a list of regular expressions and used it to collect posts that referenced movie titles. Since the focus was on US movies, we excluded posts in languages other than English; and to filter out obvious spam posts, we excluded posts that contained more than 30 links. The remaining posts effectively became our working data set from which we extracted our features.

In addition to features that counted movie references in blogs, we collected distributor, genre, and budget of each movie from `the-numbers.com`. The features that counted movie references in blogs can be classified into the following six categories:

1. Basic features that count movie references in blogs;
2. Features that count movie references taking into account ranking and indegree¹ of the blogs where they appear;
3. Features that consider only references made within a time window before or after a movie release date;
4. Features that consider only positive sentiment references;
5. Features that address spam;
6. Combinations of the above.

For the first category, we hypothesized that the importance of a movie reference in a post may be dependent on whether it occurs in the title or text of the post and thus separated title references in its own feature. On the other hand, for the second set of features, we hypothesized that the importance of a movie reference may be related to the ranking (i.e., blog popularity) and indegree (i.e., number of incoming links) of the blog where it appears. Hence, we designed features that counted references weighted by ranking ($1/\ln(rank)$) and indegree ($\ln(indegree)$) as well as

¹Both blog rank and blog indegree were calculated by `spinn3r.com`.

discretized movie references by the ranking tier of a blog in which they appear, e.g. references in blogs ranked 1-10, 11-20, 21-30, etc.

For the third set of features, our intuition was that the movie ‘buzz’ in online chatter just before and right after the release date should be indicative of movie sales and ratings. Accordingly, we discretized movie references by weeks in the following features: movie references in 5th week (35-28 days) before release date, 4th week (28-21 days) before release date, . . . , 5th week (28-35 days) after release date.

For the fourth set of features, we hypothesized that sentiment may play a role in determining the importance of movie references as predictors of movie success (as shown by (Mishne, G. et. al. 2006)). To determine sentiment, we employed a hierarchical classification approach (Pang and Lee 2004), where we used LingPipe classifiers to select from the five sentences around a movie reference ‘subjective’ ones² and to decide from these sentences sentiment polarity³. However, we found that the polarity classifier tends to be overly conservative (by assigning negative scores more frequently). Hence, we decided to include two sets of sentiment features: one associated with a conservative assessment of sentiment, where we give credit only to the posts classified positive, and one associated with an aggressive assessment of sentiment, where we give credit not only to the positively classified posts but also to the posts classified negative with a low confidence⁴.

Although we filtered a large number of spam posts during parsing by discarding those with more than 30 links, we still ended up with many spam posts in our working data set. We observed that the majority of spam posts were either very short in length (usually a sentence or two with a link) or contained many HTML tags and images. Hence, we decided to employ the following heuristic for filtering these spam posts: if a post is either less than 200 characters long or contains on average less than 20 characters of text in between the HTML tags, then it is a spam post. All of the features that relied on sentiment analysis only counted posts that satisfied this heuristic.

Finally, we designed features that consider multiple factors described above. For example, we included features such as non-spam posts with aggressive positive sentiment in blogs ranked 20–30, non-spam posts in the second week after the release date, posts with conservative positive sentiment weighted by blog in-degree, etc. For a complete list of features we used, refer to the appendix.

We found that references normalized into a 10-week time window around the release date and sales normalized into a 5-week window post release are almost perfectly propor-

²The probability of a sentence being subjective as estimated by the classifier needs to be at least 0.7.

³Both subjectivity and polarity classifiers we used were 8-gram language model classifiers, where subjectivity classifier was trained on the IMDB plot summaries and Rotten Tomatoes customer reviews and polarity classifier was trained on the full text IMDB movie reviews.

⁴We set the threshold level for aggressive sentiment to be the 0.05 cross-entropy between positive and negative sentiment probabilities.

tional to the aggregate references and sales for the entire year⁵. For this reason, total gross was an output variable and our features counted movie references in blog posts in the entire year.

Experiments and Discussion

A. Which features are most predictive of overall movie sales and rankings?

To find which features were most predictive, we filtered out movies that received less than 1000 blog references (either in the title or text of a post) and then measured Pearson’s correlation and Kullback–Leibler divergence between each feature and the variables of interest (2008 gross revenue, critics ratings and user ratings). To calculate KL divergence, we discretized our features and output variables into 20 buckets of equal size and measured divergence as per its formula for probability distributions of two discrete random variables:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

In our setting, we set P to be the distribution of an output variable and Q to be the distribution of a feature with a smoothing factor of $\lambda = 0.7$.

Correlation and KL divergence between sets of features and output variables is given in Table 1, 2, and 3. Overall, relative importance of features amongst themselves as predictors of an output variable is similar across all three variables. However, correlation of features to both critics and user ratings is low, while correlation of features to 2008 gross is very strong. For this reason, *in the rest of this paper we focus our analysis only on movie sales.*

Feature Set	Correlation	KL
References in post title	0.850	~ 0
Raw count of references in title or body	0.804	~ 0
References weighted by indegree	0.818	~ 0
References 5–3 weeks before release	0.548	0.100
References 2 weeks around release	0.822	0.010
References 3–5 weeks after release	0.847	0.114
References in top 1000 ranked blogs	0.720	0.680
References in 1000+ ranked blogs	0.672	1.030
Budget	0.665	0.991
Genre	–	1.221
Distributor	–	0.440

Table 1: Mean correlation and mean KL divergence of feature sets with 2008 gross

As seen from Table 1, which summarizes our sets of features and their correlation to gross, the most ‘predictive’ feature turned out to be the number of movie references in post titles. The correlation of references in titles to gross is 0.85 compared to correlation of references in title or text of 0.80 possibly indicating that text references are prone to false

⁵Blog references in a 10-week time window around the release date have a correlation of 0.894 with the blog references in the entire year, while sales during the first 5 weeks after the release date have a correlation of 0.951 with 2008 gross.

Feature Set	Correlation	KL
References in post title	0.095	0.060
Raw count of references in title or body	0.112	0.060
References weighted by indegree	0.100	0.060
References 5–3 weeks before release	0.017	0.022
References 2 weeks around release	0.129	0.043
References 3–5 weeks after release	0.152	0.022
References in top 1000 ranked blogs	0.095	0.588
References in 1000+ ranked blogs	0.056	0.942
Budget	0.031	0.874
Genre	–	1.156
Distributor	–	0.313

Table 2: Mean correlation and mean KL divergence of feature sets with critics ratings

Feature Set	Correlation	KL
References in post title	0.073	0.099
Raw count of references in title or body	0.097	0.099
References weighted by indegree	0.090	0.099
References 5–3 weeks before release	-0.025	0.043
References 2 weeks around release	0.121	0.083
References 3–5 weeks after release	0.134	0.034
References in top 1000 ranked blogs	0.099	0.555
References in 1000+ ranked blogs	0.072	0.887
Budget	0.026	0.841
Genre	–	1.117
Distributor	–	0.227

Table 3: Mean correlation and mean KL divergence of feature sets with user ratings

positives. Furthermore, gross revenue is highly correlated to the time series features and, in particular, references in the weeks after the release date (see Figure 1). However, references more than 3 weeks before the release date, with a low correlation of 0.55, seem to be just ‘hype’ (confirming our choice of a 5-week window). At the same time, our sentiment analysis did not improve correlation whereas ranking seems to have some importance. References in blogs ranked below 1000 have a lower correlation to the gross compared to the correlation of references in top 1000 ranked blogs, but the dependence of gross on blog ranking does not seem to be at the granular level we anticipated (e.g. the correlations for blogs ranked 1-100 and 700-800 are similar). Similarly, ac-

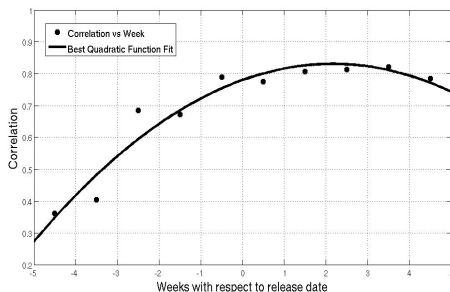


Figure 1: Time Series Feature Correlation to 2008 Gross

counting for indegree of a blog seems to improve correlation to gross but not significantly, which indicates its marginal importance. In any case, many of our features improve correlation of the raw count of references to gross which suggests that careful extraction of features is important for accurate sales prediction.

B. How do we select features for prediction of Sales?

Even though many of the features are highly correlated to the gross, before we can claim that movie sales can be predicted with blog data, we would like to see if the features themselves are similar to each other. Hence, we performed k -means clustering on the features after normalizing feature vectors using $L2$ norm. We found that over different initializations of the k -means algorithm ($k = 8..15$), we obtain consistent results, where movie references 2–4 weeks before the release date, movie references 2–4 weeks after the release date, and movie references 2 weeks around the release date were always in their own clusters. Moreover, movie references in high-ranked blogs were usually in a different cluster than references in the low-ranked blogs (1000+), and, finally, title references were consistently in the same cluster, as references in the 2–4 weeks after the release date. The latter may indicate that the global mentions of a movie is equivalent to seeing if the movie can ‘sustain’ the ‘buzz’ 2–4 weeks after its release. More importantly, this analysis shows that many of the top features are very similar, thus a good prediction algorithm, instead of picking top 15–20 most correlated features, should pick features from the distinct feature sets classified in Table 1 or apply techniques like PCA that would reduce dimensionality of data without losing much information.

C. Can we predict future sales using blog references?

Our feature analysis showed that the time series features (references in i th week before and after release date) have a very strong correlation to the movie sales. To illustrate this, consider Figure 2, where we plot normalized sales and references vs. time for the top 30 selling movies. Sales seem to follow very closely references in blogs but with a distinct lag where blog references seem to precede sales by a week. To verify this observation, we decided to do an analytical comparison of time series for sales and references similar to the one done for book sales ranks by (Gruhl, D. et. al. 2005). We used the standard cross-correlation function of two time series, as given below:

$$r_{xy}(k) = \frac{\sum_{i=1}^{n-k} (x_i - \mu_x)(y_{i+k} - \mu_y)}{\sigma_x \sigma_y}, k \in \{0, \dots, n-1\}$$

$$r_{xy}(-k) = \frac{\sum_{i=1}^{n-k} (x_{i+k} - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}, k \in \{1, \dots, n-1\}$$

where $x = x_1, \dots, x_n$, $y = y_1, \dots, y_n$ are time series of sales and references, respectively, μ_x and μ_y are their means, σ_x and σ_y are their standard deviations, and k is the lag in weeks. Using these formulas, we calculated cross-correlation at various lags for each movie and then averaged correlations across the top 30 movies to obtain an average time series correlation for each lag. As seen in Figure 3, we observed the best lag (lag where cross-correlation is the highest) at -1. A maximum cross-correlation value of 0.49

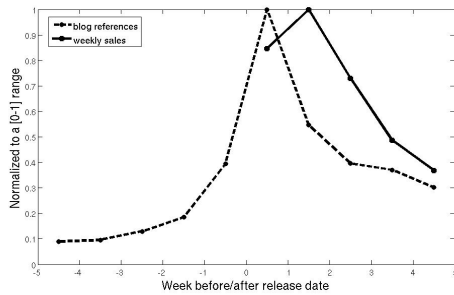


Figure 2: Normalized weekly sales and references for top 30 movies

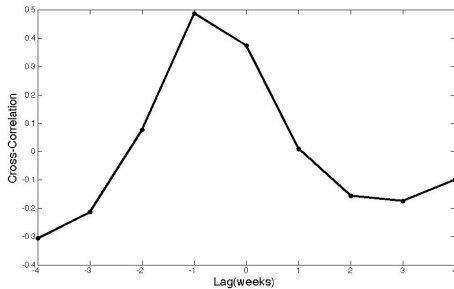


Figure 3: Cross-correlation of references and sales time series at various lags (averaged across top 30 selling movies): references precede sales by one week

at a negative lag of -1 suggests that sales tend to follow blog references in time and, hence, *references could be used as a predictor of sales, with lag of one week.*

Having confirmed our intuition that blog references precede references in time, we decided to see how feasible it would be to predict sales in week i given blog references in weeks $< i$. Hence, we measured correlation of our features to sales in week 1..5 after the release date. We found that the features, and, in particular, references before release date, have a relatively low correlation to the sales of week 1 (≈ 0.5) and budget having the highest correlation of 0.6. But for all subsequent weeks correlations get significantly better and references in blogs in prior weeks serve as good predictors of sales, with correlation values as high as 0.86.

Conclusions

We analyzed a large set of features that seem relevant in making predictions about movies. We found movie sales to be more ‘predictable’ than user ratings and critics ratings and identified the most valuable features for making such predictions (along with the analysis of similarity between the features). We found that blog references discretized by weeks around the release date are some of the most correlated features to gross and with time series analysis showed that references in blogs tend to precede movie sales. The lag between the sales and references for the top selling movies is usually one week, thus movie sales as far as a week ahead could be predicted from the blog references. Finally, we discover that sales during the first week after movie release date

may be hard to predict using blog data but become ‘more predictable’ as we move away from the release date.

Although we found high correlation of the blog features to the movie sales, prediction may still be hard. There are many features one could consider in addition to the features we looked at, e.g., season in which movie was released, number of other ‘good’ (using some metric) movies released in the same period, and other temporal events around the release date that may have an effect on sales. However, since the number of movies we are analyzing is small, increasing the number of features may not help our prediction task. On the other hand, increasing the number of movies may not be feasible either. Movies with sales less than the ones in the top 300 may not generate enough ‘buzz’ to be predictive.

Nonetheless, there are several reasons why we chose to examine the predictive potential of blogs for movie success. Firstly, movies have a known release date, which allows us to study the ‘hype’ before the release in relation to ‘success’ post release. Secondly, movies provide an inherent normalization compared to other domains, since movie sales in n th week after the release date are comparable across movies. For this reason, (Mishne, G. et. al. 2006) examines first weekend sales, while we look at the first 5-week sales post release. Finally, movie sales in the first few weeks after release tend to provide a good indication of the overall movie success (demonstrated in the high correlation between sales in the first 5 weeks and yearly gross).

The analysis we present in this paper could be generalized to other domains, such as music, consumer products, books, TV shows, etc. In fact, our findings are similar to those presented by Gruhl et. al. (Gruhl, D. et. al. 2005). However, making predictions using blog data for these domains might be harder because i) these items may not always have a strict release date, ii) the distribution of sales in time may not be comparable across items, iii) they may not generate as much ‘buzz’ in social media in order to be predictive. Hence, a different approach may be adopted, where one would track spikes in trends instead of behavior around the release date.

Acknowledgments

We would like to thank Jure Leskovec and `spinn3r.com` for providing us with the data. We also thank Hector Garcia-Molina and Andrew Ng for their insightful feedback.

References

- Gruhl, D. et. al. 2005. The predictive power of online chatter. In *KDD '05*.
- Mishne, G. et. al. 2006. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symp. on Computational Approaches to Analysing Weblogs*.
- Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 271–278.

Appendix: Full List of Features

1	mentions in the title or description	60	conservative sentiment mentions in the period between 7–14 days
2	mentions in the title only	61	(1–2 weeks) after the release date
3	mentions in the title or description of a ranked blog	62	conservative sentiment mentions in the period between 14–21 days
4	mentions in the title or description weighted by blog ranking where weights are equal to $1/\ln(\text{ranking})$ or 0 for non-ranked blogs	63	(2–3 weeks) after the release date
5	mentions in the title or description weighted by blog in-degree where weights are equal to $\ln(\text{indegree})$ or 0 for in-degree < 3	64	conservative sentiment mentions in the period between 21–28 days
6	mentions in the title or description of a non-spam blog	65	(3–4 weeks) after the release date
7	mentions in the title of a non-spam blog	66	conservative sentiment mentions in the period between 28–35 days
8	mentions in the title or description of a non-spam blog weighted by in-degree	67	(4–5 weeks) after the release date
9	mentions in the period between 35–28 days (5–4 weeks) before the release date	68	conservative sentiment mentions in the blogs ranked 1–10
10	mentions in the period between 28–21 days (4–3 weeks) before the release date	69	conservative sentiment mentions in the blogs ranked 10–20
11	mentions in the period between 21–14 days (3–2 weeks) before the release date	70	conservative sentiment mentions in the blogs ranked 20–30
12	mentions in the period between 14–7 days (2–1 weeks) before the release date	71	conservative sentiment mentions in the blogs ranked 30–40
13	mentions in the period between 7–0 days (1–0 weeks) before the release date	72	conservative sentiment mentions in the blogs ranked 40–50
14	mentions in the period between 0–7 days (0–1 weeks) after the release date	73	conservative sentiment mentions in the blogs ranked 50–60
15	mentions in the period between 7–14 days (1–2 weeks) after the release date	74	conservative sentiment mentions in the blogs ranked 60–70
16	mentions in the period between 14–21 days (2–3 weeks) after the release date	75	conservative sentiment mentions in the blogs ranked 70–80
17	mentions in the period between 21–28 days (3–4 weeks) after the release date	76	conservative sentiment mentions in the blogs ranked 80–90
18	mentions in the period between 28–35 days (4–5 weeks) after the release date	77	conservative sentiment mentions in the blogs ranked 90–100
19	mentions in the blogs ranked 1–10	78	conservative sentiment mentions in the blogs ranked 100–200
20	mentions in the blogs ranked 10–20	79	conservative sentiment mentions in the blogs ranked 200–300
21	mentions in the blogs ranked 20–30	80	conservative sentiment mentions in the blogs ranked 300–400
22	mentions in the blogs ranked 30–40	81	conservative sentiment mentions in the blogs ranked 400–500
23	mentions in the blogs ranked 40–50	82	conservative sentiment mentions in the blogs ranked 500–600
24	mentions in the blogs ranked 50–60	83	conservative sentiment mentions in the blogs ranked 600–700
25	mentions in the blogs ranked 60–70	84	conservative sentiment mentions in the blogs ranked 700–800
26	mentions in the blogs ranked 70–80	85	conservative sentiment mentions in the blogs ranked 800–900
27	mentions in the blogs ranked 80–90	86	conservative sentiment mentions in the blogs ranked 900–1000
28	mentions in the blogs ranked 90–100	87	conservative sentiment mentions in the blogs ranked 1000+
29	mentions in the blogs ranked 100–200	88	aggressive sentiment mentions in the title or description
30	mentions in the blogs ranked 200–300	89	aggressive sentiment mentions in the title only
31	mentions in the blogs ranked 300–400	90	aggressive sentiment mentions in a ranked blog
32	mentions in the blogs ranked 400–500	91	aggressive sentiment mentions weighted by blog ranking
33	mentions in the blogs ranked 500–600	92	aggressive sentiment mentions weighted by in-degree
34	mentions in the blogs ranked 600–700	93	aggressive sentiment mentions in the period between 35–28 days (5–4 weeks) before the release date
35	mentions in the blogs ranked 700–800	94	aggressive sentiment mentions in the period between 28–21 days (4–3 weeks) before the release date
36	mentions in the blogs ranked 800–900	95	aggressive sentiment mentions in the period between 21–14 days (3–2 weeks) before the release date
37	mentions in the blogs ranked 900–1000	96	aggressive sentiment mentions in the period between 14–7 days (2–1 weeks) before the release date
38	mentions in the blogs ranked 1000+	97	aggressive sentiment mentions in the period between 7–0 days (1–0 weeks) before the release date
39	non-spam mentions in the period between 35–28 days (5–4 weeks) before the release date	98	aggressive sentiment mentions in the period between 0–7 days (0–1 weeks) after the release date
40	non-spam mentions in the period between 28–21 days (4–3 weeks) before the release date	99	aggressive sentiment mentions in the period between 7–14 days (1–2 weeks) after the release date
41	non-spam mentions in the period between 21–14 days (3–2 weeks) before the release date	100	aggressive sentiment mentions in the period between 14–21 days (2–3 weeks) after the release date
42	non-spam mentions in the period between 14–7 days (2–1 weeks) before the release date	101	aggressive sentiment mentions in the period between 21–28 days (3–4 weeks) after the release date
43	non-spam mentions in the period between 7–0 days (1–0 weeks) before the release date	102	aggressive sentiment mentions in the period between 28–35 days (4–5 weeks) after the release date
44	non-spam mentions in the period between 0–7 days (0–1 weeks) after the release date	103	aggressive sentiment mentions in the blogs ranked 1–10
45	non-spam mentions in the period between 7–14 days (1–2 weeks) after the release date	104	aggressive sentiment mentions in the blogs ranked 10–20
46	non-spam mentions in the period between 14–21 days (2–3 weeks) after the release date	105	aggressive sentiment mentions in the blogs ranked 20–30
47	non-spam mentions in the period between 21–28 days (3–4 weeks) after the release date	106	aggressive sentiment mentions in the blogs ranked 30–40
48	non-spam mentions in the period between 28–35 days (4–5 weeks) after the release date	107	aggressive sentiment mentions in the blogs ranked 40–50
49	conservative sentiment mentions in the title or description	108	aggressive sentiment mentions in the blogs ranked 50–60
50	conservative sentiment mentions in the title only	109	aggressive sentiment mentions in the blogs ranked 60–70
51	conservative sentiment mentions in a ranked blog	110	aggressive sentiment mentions in the blogs ranked 70–80
52	conservative sentiment mentions weighted by blog ranking	111	aggressive sentiment mentions in the blogs ranked 80–90
53	conservative sentiment mentions weighted by in-degree	112	aggressive sentiment mentions in the blogs ranked 90–100
54	conservative sentiment mentions in the period between 35–28 days (5–4 weeks) before the release date	113	aggressive sentiment mentions in the blogs ranked 100–200
55	conservative sentiment mentions in the period between 28–21 days (4–3 weeks) before the release date	114	aggressive sentiment mentions in the blogs ranked 200–300
56	conservative sentiment mentions in the period between 21–14 days (3–2 weeks) before the release date	115	aggressive sentiment mentions in the blogs ranked 300–400
57	conservative sentiment mentions in the period between 14–7 days (2–1 weeks) before the release date	116	aggressive sentiment mentions in the blogs ranked 400–500
58	conservative sentiment mentions in the period between 7–0 days (1–0 weeks) before the release date	117	aggressive sentiment mentions in the blogs ranked 500–600
59	conservative sentiment mentions in the period between 0–7 days (0–1 weeks) after the release date	118	aggressive sentiment mentions in the blogs ranked 600–700
		119	aggressive sentiment mentions in the blogs ranked 700–800
		120	aggressive sentiment mentions in the blogs ranked 800–900
		121	aggressive sentiment mentions in the blogs ranked 900–1000
		122	aggressive sentiment mentions in the blogs ranked 1000+
			distributor
			genre
			budget