Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2009) 000–000

Procedia Social and Behavioral Sciences

www.elsevier.com/locate/procedia

COINs2009

Predicting Movie Prices Through Dynamic Social Network Analysis

Lyric Doshi[a], Jonas Krauss[abc], Stefan Nann[abc], Peter Gloor[a]*

[a]*MIT Center for Collective Intelligence*
[b]*University of Applied Sciences Northwestern Switzerland*
[c]*University of Cologne*

Elsevier use only: Received date here; revised date here; accepted date here

Abstract

This paper explores the effectiveness of social network analysis and sentiment analysis in predicting trends. Our research focuses on predicting the success of new movies over their first four weeks in the box office after opening. Specifically, we try to predict prices on the Hollywood Stock Exchange (HSX), a prediction market on movie gross income, and predict the ratio of gross income to production budget.

When predicting movie stock values on HSX, we consider two different approaches. One approach is to predict the daily changes in prices. This means we would be predicting a mix of not only how well we think the movie will perform, but also how we think other people think the movie will perform. Our second approach is to predict the final closing price of the stock, which will be how much the movie actually grosses in the box office after four weeks. In this approach, the daily prices provide feedback with the crowd's constantly revising estimate of the final performance of the movie. Finally, we try to classify movies in three groups depending on whether they gross less than, just over, or a lot more than their production cost.

For our prediction we gather three types of metrics. (1) Web Metrics are movie-rating metrics from IMDb and Rotten Tomatoes as well as box office performance data from Box Office Mojo and movie quotes from HSX. (2) SNA Metrics Web and blog betweenness represent the general buzz on the movie from the web and from bloggers. We hypothesize that they will be useful because they are unconscious signals about a movie's popularity. (3) To determine the general sentiment about the movies, we gather posts from IMDb forums to generate Sentiment Metrics for positivity and negativity based on the discussion in the forums. Our preliminary results employing different prediction methods such as multilinear and non-linear regression combining our three types of independent variables are encouraging, as we have been able to predict final box office return at least as good as the participants in the HSX prediction market.

Keywords: Dynamic social network analysis, trend prediction, movie analysis

1. Introduction

The purpose of this research is to explore the effectiveness of social network analysis and sentiment analysis in predicting trends by mining publicly available online data sources. The high-level hypothesis of the project is that the collective whole can provide valuable predictive information about events such as elections and stock market results. We aim to garner the opinions of the many on the Web to make meaningful predictions by putting together the little pieces of the big picture each person can provide. More specifically, this research focuses on

predicting the success of new movies over their first four weeks in the box office after opening as a medium for testing this hypothesis. Our aim is to predict prices and trade on the Hollywood Stock Exchange (HSX), a prediction market where each movie stock price represents people's anticipation of the movie's box office success. We also hope to correctly identify movies as potential flops or blockbusters based on how much the gross relative to their production budget soon after they open in the box office.

2. Method

When predicting movie stock values on HSX, we consider two different approaches. The first approach is to predict the daily changes in prices. This means we would be predicting a mix of not only how well we think the movie will perform, but also how we think other people think the movie will perform. This is a slightly different problem than the second approach, which is to predict the final closing price of the stock that will be how much the movie actually grosses in the box office after four weeks. In the second approach, the daily prices provide feedback with the crowd's constantly revising estimate of the final performance of the movie.

After some experiments with the first approach, we have opted to pursue the second approach of predicting the final success of the movie rather than guess its daily changes. The expected variance on predicting the final value should be much less than the expected variable of daily changes because longer term trends tend to be more stable. Daily price changes are subject to sudden sways in opinion and the fickleness of traders.

2.1. Data Sources

To predict HSX prices, we gather many raw and derived independent variables. We categorize them as either Web Metrics, Social Network Analysis Metrics, or Sentiment Metrics. We also provide a justification of how these metrics help us capture the wisdom of the collective whole.

2.2. User Ratings

We gather movie rating metrics from IMDb (www.IMDb.com) and Rotten Tomatoes (www.rottentomatoes.com) as well as box office performance data from Box Office Mojo (www.boxofficemojo.com). The movie quote prices themselves are gathered from the Hollywood Stock Exchange site (www.hsx.com).

IMDb aggregates votes over a large number of users. Anyone can submit their rating of a movie to be included in the IMDb rating. Thus, IMDb provides a summary of the overall opinion of the collective whole and we hypothesize represents the general feeling about the quality of a movie. However, since the voting is open, the rating is also susceptible to users trying to bias the vote artificially.

Rotten Tomatoes, on the other hand, collects the input of movie critics. We view this as an aggregation of the opinions of movie "experts" only. The number of contributors is smaller than IMDb, but in theory each vote may provide a better quality input into the overall vote. Here we are polling the collective "expert" mind, but again we may find this vote susceptible to critic bias.

Both rating sites may also produce snow ball effects, especially on IMDb, where strong positive or negative reviews encourage more people to see or not see the movie. We hope to try to capture this behavior in our models if it does exist.

The Box Office Mojo provides day to day data on how much each movie has grossed in the box office to date. The income provides a reflection of the collective whole's general opinion regarding a movie. Initial high income that peters out represents a highly hyped movie that failed to live up to its reputation. Sustained income may represent the expected success of a good movie. Alternatively, a ramp in income over time represents a Black Swan such Slumdog Millionaire that gains popularity as more people see and praise it. It also provides an empirical corrective factor for the final HSX price we are trying to predict.

2.3. Social Network Analysis Metrics

The first step to measuring a trend with dynamic social network analysis is the tracking of a movie title's relative importance on the Web, in the blogosphere, or in an online forum. We call each of these different online communication archives an information sphere. As an approximation for the relative importance of a concept in the information sphere, we calculate the betweenness centrality of this concept within the chosen information sphere. This means that we are extending the well-known concept of betweenness centrality of actors in social networks to semantic networks of concepts – movie titles in this case.

The betweenness centrality of a concept in a social network is an approximation of its influence on the discussion in general. Betweenness centrality in social network analysis tracks the number of geodesic paths through the entire network, which pass through the concept whose influence is measured. As access to knowledge and information flow are means to gain and hold on to power, the betweenness centrality of a concept within its semantic network is a direct indicator of its influence (Wassermann & Faust 1995). In other words, concepts of high betweenness centrality are acting as gatekeepers between different domains. While communication in online forums can be used to construct social networks among actors, we can also construct social networks from blogs and the Web. Although these semantic networks based on blog and Web links are not true social networks in the original sense, they are straightforward to construct by considering the Websites and blog posts as nodes and the links between the Websites and blog posts as ties of the social network.

Measuring the betweenness centrality of a concept permits us to track the importance of a concept on the Web or in the Blogosphere. This can be done either as a one-time measurement, or continuously in regular intervals over time, as Web pages, blog posts, and forum posts all have time stamps. We therefore periodically (e.g. once per day, once per hour, etc.) calculate the betweenness centrality of the concept. The resulting betweenness centrality is a numerical value between zero and one, with zero implying no importance of the concept in the information sphere and values above zero representing the relative importance in comparison to other concepts.

To build the semantic social network in an information sphere we introduce degree-of-separation search. Degree-of-separation search works by building a two-mode network map displaying the linking structure of a list of Web sites or blog posts returned in response to a search query, or the links among posters responding to an original post in an online forum. For example, a search to get the betweenness of "Hillary Clinton" on the Web works as follows:

1) Start by entering the search string "Hillary Clinton" into a search engine.

2) Take the top N (N is a small number, for example 10), of Web sites returned to query "Hillary Clinton".
3) Get the top N Web sites pointing to each of the returned Web sites in step 2 by executing a "link:URL" query, where URL is one of the top N Web sites returned in step 2. The "link:" query returns what the search engine considers "significant" Web sites linking back to a specific URL.
4) Get the top N Web sites pointing to each of the returned Web sites in step 3. Repeat step 4 up to the desired degree of separation from the original top N Web sites collected in step 2. Usually it is sufficient, however, to run step 4 just once.

Degree-of-separation search therefore is very similar to a domain-specific page-rank algorithm (Brin & Page 1998).

The betweenness metrics represent the general buzz on the movie from the web and from bloggers. We hypothesize that they will be useful variables because they are unconscious signals about a movie's popularity (or notoriety). That is, they are not calculated by active input from people and are therefore difficult to influence artificially.

2.4. Sentiment Analysis Metrics

To determine the general sentiment about the movies, we gather posts from IMDb forums. We previously used the following general forums: Oscar Buzz, Film General, Box Office, and Now Playing and Upcoming Films. However, we are also tracking communication on movie-specific forums to allow us to better differentiate which posts are about which movie. We are currently investigating a few different sentiment analysis algorithms, described in the following section, to generate a score for positivity and negativity. When using general forums, we also counted the occurrences of the movie's title.

In addition to calculating a sentiment score for each post, we also build a social network of all the post authors in order to calculate their betweennesses. We can then weight the post score by the betweenness of its author. This gives the posts by more between, and we hypothesize more influential, authors relatively higher sentiment scores.

2.4.1. Basic Sentiment Algorithm

Pang and Lee, 2002, have shown that automatic extraction of words and word pairs leads to more precise results than manually selecting positive and negative words. Our approach follows the basic "bag-of-words" approach which considers the co-occurrences of keywords in sentences or text (Nasukawa et al, 1999). A drawback of this approach is the disregard of grammatical dependencies in the analyzed data. This might lead to misleading interpretation in some cases. For example the statement "Terminator is not good" would be classified as a positive sentiment with the simple "bag-of-words" approach. In practice this problem seems to be rare, however. (Matsuzawa & Fukuda, 2000) state that 40% of analyzed keywords in the same sentence or text block show grammatical dependencies. By reading a large sample of forum messages we empirically verified their finding that actors mostly use negative phrases rather than negating positive phrases when they want to express something negative. For example they use the phrase "is bad" instead of "is not good". We further reduce occurrence of this problem through not looking at the whole post but rather only words around a word anchor.

Since discussion about different movies involves different words with occasionally conflicting positive and negative sentiment expression, we hypothesize that adapting the bag-of-words to specific types of movies will deliver better results than one general bag-of-words for all kinds of

movies. For example, comedies are deemed positive if they make people laugh, however, horror movies are supposed to scare people. Thus, in the first case words like "laugh", "hilarious", and "funny" indicate a positive sentiment, while in the second case these words would imply the opposite. We aim to dynamically adapt the bag-of-words for the analysis of movies and suggest three alternative methods.

The starting point of the sentiment retrieval is the collection of word lists that constitute the initial bag-of-words. These lists were retrieved from the movie discussion on the IMDb forum and manually checked to assess the word's appropriateness. One list is used for positive words, one for negative. To deal with different cases, singular and plural forms, etc., we apply Porter Stemming (Rijsbergen et al. 1980). Through the application of stop lists to forum posts we sort out unimportant words like "the", "and", "or", etc.

When analyzing IMDb posts, we only consider the individual forum that discusses the current movie of interest. Each movie has its own message board which is being used as its document corpus. These words form the basis for the comparison with the bag-of-words. Generally, when more words from the positive list are present, the positive sentiment value is higher. The same is true for negative words and negative sentiment value.

Using the adapted bag-of-words, we examine the ratio of positive to negative words in IMDb discussion about a particular movie. The movie's genre tells which bag-of-words has to be used (e.g. comedy bag-of-words for a comedy). Posts of different forum actors are weighted differently in their contribution to the ratio based on the betweenness centrality of the actor (the higher the betweenness, the higher the weight). Since the sizes of the positive and negative bag-of-words vary, words being added are weighted by the number of words in the analyzed post divided by the number of words in the applied bag-of-words, based on the assumption that each word from the bag-of-words has the same likelihood to appear in the discussion. The ratio is calculated for single days, providing a sentiment value for prediction purposes. Generally a day with a higher positive to negative ratio should indicate a quote rise for this movie on HSX and vice versa.

2.4.2. Dynamic Adaptation of Bag-of-Words

Three different ways of adding new words to the existing word lists in the bag-or-words are currently being evaluated. The common basic principle is the initial determination whether a post is positive or negative. The approaches differ in the way of how they determine this property. The first method is to base a post's sentiment property on the current trend in the HSX prediction market. The second method derives the sentiment property of a post from the general reputation of a movie as expressed by critics' ratings. In the third method, it can be determined through the ratio of positive to negative words surrounding each prospective new addition to the bag-of-words in a post.

Once a post's sentiment property is captured, its content is added to a continuously updated word list. Multiple entries are summed up and thus ranked higher. After processing all posts about a specific movie we retrieve the most frequent words and add them to the positive or negative bag-of-words for that movie, depending on the previously determined positivity / negativity property of the post. The other words from the list are discarded. The result of this process is a bag-of-words each for positivity and negativity for a particular movie genre. The quality of the bag-of-words can be increased through repeating the process for multiple movies of the same genre.

This approach is different from other methods as support vector machines, naïve Bayes, or maximum entropy classification. It does not require manual review of posts to tag them as positive or negative, making the system completely automatic. However, since the training set determines the quality of the dynamic bag-of-words, it is important to have a high ratio of correctly classified post at the beginning of the process.

3. Modeling

Once we have our collected and derived data, we wish to develop a model to use some subset of our independent variables to predict future HSX prices. In this section, we discuss our current results and future plans. We have considered five different modeling approaches, which will subsequently be described.

3.1. Direction Prediction

We briefly investigated simply predicting whether the price tomorrow will go up or down based on the changes in variables on previous days. In the simplest case, we qualitatively tested the hypothesis that on day j, the sign of the delta of some independent variable between day j-1 and j would match the delta of the price between day j and j+1. Expanding the idea from the delta, we instead used the slope of the regression line on the series day j-modelingDays to j against the independent variable series from day j-modelingDays to day j to predict the change in price between day j and day j+1. Along with considering each independent variable against the price, we also pooled variables with changes that matched the HSX prices to make consensus votes on the direction the price would change. For example, on day j, we could perhaps have each of the three most accurate independent variables so far contribute a vote of whether the price will go up or down tomorrow. The majority vote becomes the prediction.

The coarse-grained nature of this approach, however, leaves it susceptible to coincidental correlations. Furthermore, simply knowing if a price will go up and down is not very useful without having a guess at the magnitude of the change. Without magnitude, it is difficult to make meaningful decisions about which movie stocks to buy or sell each day. Also, the coarse and local nature of this type of prediction does not allow is to guess at the actual final stock value. That said, the qualitative experiments did show some correlation between price changes and variables like web and blog betweenness, sentiment, and gross revenue. Over the course of many different trials, these variables were accurate enough to be used in consensus voting over twice as often as the other independent variables. While not necessarily useful directly, these results give us some confidence in these variables being good indicators when we try to predict the final stock price using different methods.

3.2. Linear Regression

For our first attempt to predict the magnitude of the movie stock price, we used linear regression to predict the price tomorrow on day j+1. We based our predictions on a model slope and intercept built using previous independent variable and movie stock price data. Over many trials, we varied the modelingDays, or the number of days of data used to build the model, and the predictionDays, or the number of days the price was shifted forward for each independent variable and price value pairing. To clarify, to test the hypothesis that day i's independent variable value affects day i+1's price and that the previous 5 days data best helps us predict the pattern for tomorrow, predictionDays would be set to 1 and modelingDays would be set to 5. To calculate the prediction for day 7, we would use the independent variable from days 1-5 and the

prices from days 2-6 to build a linear regression model. Then we enter the independent variable value at day 6 as input to compute the predicted price output for day 7. Increasing predictionDays to 2 would mean that the independent variable from days 0-4 would be used with the prices from days 2-6 to build the model. Thus, predictionDays represents the lag in the effects of the independent variable on the price, while modelingDays encapsulates the number of the days that define a short term trend to model in the price. Especially when considering movies, it is very likely that there will be a lag between a positive or negative set of reviews posted online regarding a movie and the subsequent boosting or damaging effect on its price.

We tried many combinations of values for both variables over all movies to analyze how much the price changes lagged in changes to each independent variable and how many days were typically required to build a model to capture the current trend in the price. For each value of modelingDays ranging from the minimum 2 days to 14 and predictionDays ranging from the minimum 1 day to 14, we cycled through all our historical data for a movie and built a model for every sliding window of modelingDays to make a prediction. This means for modelingDays = 5 and predictionDays = 2, we'd make a total of 30 predictions if we had 37 days of data.

To evaluate each set of modelingDays and predictionDays variable values, we compared the predictions against the actual prices for those days. Specifically, we computed the mean error, standard deviation of error, and mean squared error. Since each prediction also has a correlation, we also computed the mean correlation and correlation standard deviation. Some the results here have reasonably low error, indicating we may be able to predict day to day changes reasonably to make trades. Table 1 shows results for the movie Up. The number of theaters showing the movie each day was the best predictor for the next day's movie stock price with an 8 day delay representing the lag between people seeing the movie and reacting to it on the prediction market. A mean squared error of 10.387 accumulated over 10 predictions in this case seems reasonable. It should be noted that the mean squared error is calculated using the error between each predicted price and actual price for each day. The correlation coefficient mean is an average of the correlation coefficients of all the regression models used for the predictions. Thus, the accuracy of the predictions and the correlation coefficients in this table are related, but not directly linked. We have not yet developed any trading strategies to apply these results.

IV Name	*Modeling-Days*	*Prediction-Days*	Mean Error	Error Standard Deviation	Mean Squared Error	Correlation Coefficient Mean	Correlation Coefficient Standard Deviation
Number of theaters showing film	8	2	0.006315	3.380257	10.387	0.765246	0.1445
IMDb votes for score of six	2	5	0.136708	3.348881	10.533	0.428571	0.937614
Percentage of IMDb votes for score of three	8	14	-1.366	3.134408	10.989	-0.46962	0.521188
IMDb mean score	3	11	-1.575	3.155722	11.609	-0.652	0.500621
IMDb votes for score of five	2	4	-0.2703	3.530512	11.759	0.428571	0.937614
Percentage of IMDb votes for score of three	9	14	-1.93602	2.972594	11.905	-0.50907	0.519679
IMDb weighted average score	3	12	0.1785	3.685474	12.709	-0.29391	0.756806
IMDb votes for score of six	2	8	-0.97831	3.556334	12.761	-0.07692	1.037749
IMDb mean score	4	11	-1.52655	3.359891	12.813	-0.63204	0.356732
IMDb votes for score of five	2	8	-0.17524	3.706572	12.853	0.384615	0.960769

Table 1: Top 10 linear regression prediction results for the movie Up sorted by mean squared error

Given our goal of trying to predict the final HSX closing price, which is representing the movie's box office performance, we wish to investigate another way of applying regression to our movie data. In the previous discussion, we essentially developed a regression model that worked best for individual movies. As an alternative approach, we wish to use historical data with full knowledge of "past" and "future" to develop a single model to predict movie stock prices. Given n+1 days of data on a movie ending with the movie's closing price, we build a linear regression model using the independent variable from day 0 to day k-1 matched with the prices from day n-k to day n-1. We can then use day k as input to the model to predict day n and calculate our squared error. We repeat the process for values of k from 2 to n to determine how quickly the model converges to an acceptable prediction of the price on day n, for example within 5% or 10%. The lower the value of k for which the model converges, the earlier we can make our prediction of the final movie stock price. To expand the search space, we need not only use day 0 to day k-1 to build the model for each value of k. Instead, we may also test any period of consecutive days ending in day k-1. After computing the model coefficients and ideal k for each independent variable for each movie using historical data, we will combine them into a single model that can predict a final movie stock price given a new movie's initial independent variable values ideally well before the last day stock is traded. The historical data will be used to simulate new movie data to evaluate the accuracy of the model.

3.3. Multilinear and Non-Linear Regression

The linear regression approaches provide a basic foundation, but we suspect that a combination of several independent variables together will produce the best prediction results. Realistically, we believe multiple variables should all contribute some useful information toward determining a movie's imminent success. For example web buzz gives the prominence of a movie, but sentiment clarifies whether that buzz is positive buzz or negative buzz.

We will apply parallel approaches as with linear regression, except using all the independent variables together to build a single model. L1 regularization or a similar approach will be used to identify which of the many independent variables provide significant contributions to the prediction model.

Depending on the quality of the multilinear regression, we may also try similar approaches with non-linear models. At this point, we do not know the type of model that will best relate the independent variables to the price.

3.4. Slope of Smoothed SNA and Sentiment Variables

We believe that SNA and sentiment variables should provide especially meaningful insights about the price of a movie. Initially, we wish to smooth their values with an n-day moving average. Since weekends are popular times for people to react to others' opinion and go watch movies or voice their own opinions after watching a move, we suspect a 7-day moving average will be effective. Once smoothed, the independent variable values will be used to make a model following the same approach as the linear regression modelingDays and predictionDays. For this approach, however, we propose simply using the slope of the regression either as a prediction for the actual value of the delta in price. If the slope is not the same as the delta, we propose that it may be highly correlated. This is another approach we need to test against our historical data that we've collected so far.

3.5. Classifying Movies as Successes or Flops Based on Gross-to-Production-Budget Ratios

In our final model so far we categorize the success of movies based on the ratio of their gross income to production budget using three groups. Group I movies have a ratio of less than 1. These movies are flops that do not even recoup their investment. Group II comprises movies with a ratio between 1 and 2, i.e. movies that did decently. Group III holds the most successful blockbuster movies making at least double what was invested.

We first classified each of the 30 movies with daily betweenness and sentiment values two weeks before release through a week after release into Groups I, II, and III. Then we calculated the product of blog betweenness, or buzz, and positive and negative sentiment value for each day and summed the products together. Here positive sentiment values were positive numbers and negative sentiment values were negative numbers. The net sum represents the overall feeling about the movie a week after release. We then take the absolute value and log of the sums to turn them into strength-of-feeling (whether positive or negative) scores of comparable scale. We hypothesize that these scores are normally distributed within each category. Using a leave-one-out testing policy, we cycle through all 30 data points (i.e. movies) and leave one out at a time for checking. With the remaining points, we compute the mean and standard deviation of the scores for each category. Recall by Bayes rule, if g is the group for data point x, then $P(g|x) = P(x|g)P(g)/P(x)$, which is proportional to $P(x|g)P(g)$ since $P(x)$ doesn't depend on g. At this point,

we do not have substantial data to suppose our prior P(g) is not uniform between Groups I, II, and III. Later we may be able to refine this. Thus, P($g|x$) is proportional to P($x|g$), which we can calculate using the normal distribution probability density function for which x, the group mean, and the group standard deviation are inputs. We attribute x to whichever group gives the highest value for P($x|g$).

Using this approach, we initially found the classifications to be correct 53% of the time, which is better than 33% for random guessing. Table 2 shows how often movies in each group were categorized correctly, where Group III did quite well. More usefully, Table 3 shows how often a movie classified as being in a group was actually in that group. It should be noted that Group II is at a 100%, but only 2 of 12 movies were identified as being in Group II. Finally, Table 4 shows perhaps the strongest result: That a movie classified as Group III was really unlikely to turn out to be in Group I and a movie identified as in Group I was really unlikely to be in Group III. Thus a flop suspect identified by our approach rarely made it big and a movie we named as a potential blockbuster rarely flopped.

Group	Percentage of Movies Classified Correctly Into This Group
Group I	72.7%
Group II	16.7%
Group III	85.7%

Table 2: Success rates of classifying movies into the correct groups

Group Movie was Classified as Being In	Percentage of Those Movies Actually In That Group
Group I	56.5%
Group II	100%
Group III	55.4%

Table 3: These percentages compare the number of movies classified correctly against false positives.

Percentage of Movies Classified as Group III, But Actually In Group II or III	82.4%
Percentage of Movies Classified as Group I, But Actually In Group I or II	88.9%

Table 4: This table shows that movies identified as successes usually did ok or quite well, but rarely actually ended up doing poorly. Likewise, movies classified as flops were either flops or did ok, but rarely became successes.

Improving the scores by advancing the sentiment algorithms and by testing different combinations of the metrics is one route improving this method further. Some movies have some pretty extreme sentiment ratings and this may in part be due to IMDb forum parsing irregularities. Working to grow the data set from 30 movies will also add more strength to the findings as well. Using the hype around a movie to adjust its prior may also help us do better.

4. Conclusion

In this project, our success will be measured by how accurately we can predict daily price changes or the final movie stock price. We also wish to maximize how far we can effectively predict into the future. For classifying movies as flops and blockbusters, we already do better than guessing and are fairly robust to false positives. In this paper, we've laid a foundation of classifying movies as flops- and blockbusters-to-be a week after release and being either correct or close to correct (meaning the movie was just so-so) over 80% of the time. We rarely identified a flop as a blockbuster or vice versa. We will continue to strengthen our ability to distinguish movies in each grouper by increasing our data set, improving our sentiment metrics, and exploring ways to further distinguish our groups from each other.

References

Brin, S. Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine, In Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia: Elsevier.

Das, S. R., Chen, M. Y. (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, Management Science Vol. 53 Issue 9: p 1375-1388.

Gloor, P. Krauss, J. Nann, S, Fischbach, K. Schoder, D. (2009) Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. IEEE Conference on Social Computing (SocialCom-09), Aug 29-31, Vancouver.

Gloor, P., Zhao, Y., (2006) Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis, Proceedings of 10th IEEE International Conference on Information Visualisation IV06 (London, UK, 5-7 July 2006)

Gloor, P., Zhao, Y. (2004) TeCFlow - A Temporal Communication Flow Visualizer for Social Networks Analysis, ACM CSCW Workshop on Social Networks (ACM CSCW Conference, Chicago, 6. Nov. 2004).

Matsuzawa, H.; Fukuda, T. (2000). "Mining Structured Association Patterns from Databases," Proceedings of the 4th Pacific and Asia International Conference on Knowledge Discovery and Data Mining (2000), pp. 233-244.

Nasukawa, T., Morohashi, M., Nagano, T. (1999) Customer claim mining: Discovering knowledge in vast amounts of textual data. Technical report, IBM Research, Japan, 1999.

Pang, B., Lee, L. (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. Association for Computational Linguistics. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

C.J. van Rijsbergen, S.E. Robertson and M.F. Porter, 1980. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587)

Wasserman, S., Faust, K. (1994) Social Network Analysis, Cambridge University Press.

Wysocki, P. D. (1999) "Cheap Talk on the Web: Determinants of Posting on Stock Message Boards". University of Michigan: Working Paper, November 1999.

Zitzewitz, E. Wolfers, J. (2004) "Prediction Markets", Journal of Economic Perspectives, Winter 2004