Web Science 2.0: Identifying Trends through Semantic Social Network Analysis

Web Science 2.0: Identifying Trends through Semantic Social Network Analysis

Peter A. Gloor
Center for Collective Intelligence
MIT, Cambridge, MA, USA
pgloor@mit.edu

Jonas Krauss, Stefan Nann
University of Applied Sciences
Northwest Switzerland Brugg
{jonas.krauss, stefan.nann}
@fhnw.ch

Kai Fischbach, Detlef Schoder
Dept. of Information Systems
University of Cologne
{fischbach, schoder}@wim.uni-koeln.de

Abstract—**We introduce a novel set of social network analysis based algorithms for mining the Web, blogs, and online forums to identify trends and find the people launching these new trends. These algorithms have been implemented in Condor, a software system for predictive search and analysis of the Web and especially social networks. Algorithms include the temporal computation of network centrality measures, the visualization of social networks as Cybermaps, a semantic process of mining and analyzing large amounts of text based on social network analysis, and sentiment analysis and information filtering methods. The temporal calculation of betweenness of concepts permits to extract and predict long-term trends on the popularity of relevant concepts such as brands, movies, and politicians. We illustrate our approach by qualitatively comparing Web buzz and our Web betweenness for the 2008 US presidential elections, as well as correlating the Web buzz index with share prices.**

Social network analysis, semantic social network analysis, trend prediction, Web mining

I. INTRODUCTION

The Internet has become a major communication channel for late-breaking news and to disclose innermost secrets. For example, when CBS published documents about George W. Bush's behavior during his military service, Republican bloggers quickly identified weak spots in the authenticity of the documents. This questionable evidence regarding George Bush's potential evasion of military service during the Vietnam War era ultimately lead to the early retirement of CBS news anchor Dan Rather. This incident is just one of many illustrating that today's news are made and disseminated on the Web and in the blogosphere. The Web therefore has become both part of and a mirror of the "real world". Assuming that people will be doing what they announce, analyzing what influential people say on the Web might identify trends before they have been recognized by the rest of the world [15]. Towards this goal, we introduce a new way of measuring the changes in popularity of brand names and famous people such as movie stars, politicians, and business executives, based upon the premise that in today's Internet economy, buzz on the Web reflects popularity and buzz in the real world.

The approach described in this paper mines and analyzes unstructured communication and information from Web resources. As input for our method we take concepts in the form of representative phrases from a particular domain – for example names of politicians, brands, or issues of general interest. In a first step the geodesic distribution of the concept in its communication network is determined by calculating the temporal betweenness centrality of the linking structure. The second step adds the social network position of the concept's originator – called "actor" in social network language – to the metric to include context-specific properties of nodes in the social network. In the third step we qualitatively evaluate the concept's communication context to assess the concept's perception on the Web, blog, or online forum.

Result of this three-step process is a "Web buzz index" for a specific concept that allows for an outlook on how the popularity of the concept might develop in the future. In the remainder of this paper, after an overview of the state of the art, we introduce our three-step process. We illustrate it first by tracking the presidential elections, and then by showing the correlation between fluctuations in the Web buzz index for stock titles and stock prices.

II. RELATED WORK

Popularized by Barabási [4] in his book "Linked", there is a rich body of research on how the linking structure of the Web influences accessibility of Web pages and their ranking in search engines.

Visualization of Web structure and contents has been an active area of research since the creation of the Web. There are numerous systems for the static visualization and analysis of the link structure of the Web [9], [10]. Inxight, Visual Insight, Touchgraph, Grokster, and Mooter are all systems for the visualization of the linking structure of the Web, sometimes also offering a visual front end for search results.

In a related stream of work, researchers have been trying to predict the hidden linking structure based on known links [1], [2]. Additionally, by looking at contents of Web sites, subspaces of the Web have been clustered by topics [6]. Combining these two lines of research, community Web sites have been mined to discover trends and trendsetters for viral marketing [18].

Our research focuses on a similar application – tracking the strengths of concepts over time. For our analysis we are

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.186

using the Condor system [13] originally developed to mine e-mail networks to automatically generate dynamic social network movies.

There are various studies that are dealing with the prognosis of stock prices through an analysis of online communication in blogs and message boards [22], [3]. Researchers are also basing their studies on the most popular finance-related online communities Yahoo! Finance, Raging Bull, and Motley Fool [14]. References [8], [21], [22] are applying sentiment extraction algorithms on finance-related communication data from message boards [24].

III. Categorization of Web Sources in Information Spheres

For our Web mining approach we classify the World Wide Web into three categories or information spheres: The Web at large – we call it "Wisdom of Crowds", the blogosphere – "Wisdom of Experts", and forums – "Wisdom of Swarms". Each of these three sources is processed differently in our method based on the way how the information contained in it is produced. Online forums contain the most focused and up-to-date information about a certain subject. These forums are self-organized communities consisting of individuals as well as organizational institutions, which exchange ideas and information [20]. The huge "swarms" of people in the forum represent the collective opinion of those who care most about the forum's topic.

Blogs represent the "Wisdom of Experts". The number of bloggers and new blogs grew exponentially over the last few years and is still growing. Contrarily to forums, where posters engage in a dialogue amongst themselves, bloggers are individual experts where each of them is expressing his or her private opinion. Because an expert is not always right, it would be risky to rely on a single opinion. But combining the wisdom of experts about a subject will lead to an aggregated indicator of the collective opinion of experts about a certain topic.

Finally, mining the Web at large also gives valuable clues about a certain topic. The topics might be discussed on sites of varying popularity and actuality such as online news sites, company Websites, information Websites, etc. This resource is by far the largest of the three and incorporates the collective opinion of a large part of the Western world – what we call the "Wisdom of the Crowds."

These three different data sources represent the basis for our combined communication and information analysis process.

IV. Concept Weighting Strategy

For the last six years we have developed a sophisticated semantic social network analysis tool called Condor [12], [13]. Condor (formerly called "TeCFlow") includes automated textual analysis functionality using standard information retrieval algorithms like "term frequency–inverse document frequency" [19]. Additionally, Condor factors in the betweenness centrality of actors for weighing the content by the social network position of actors.

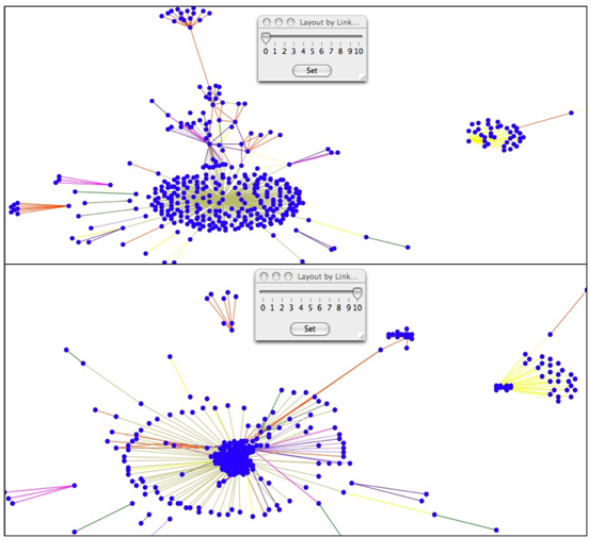

Figure 1. Weighting a set of documents by social network position of actors only (top), and also factoring in similarity of contents. All networks in this and subsequent figures are visualized with the Fruchterman-Rheingold graph layout algorithm [11].

Fig. 1 illustrates this concept by showing two Condor screen shots of the same document network. The top of the picture shows a social network of actors based on exchange of e-mails. While senders and receivers of e-mails are represented by nodes, the edges reflect an exchange of e-mails between two actors. The bottom of the picture shows the same network, but now the actors have additionally been grouped by the similarity of contents of their discussion. The blue and very dense cluster in the middle of the network represents all actors that are talking about the same subject in their e-mail communication. Clustering of nodes at the bottom of fig. 1 is therefore done by combining two attractive forces, first based on the number of exchanged e-mails, and second based on the similarity between two e-mail text bodies calculated by "term frequency–inverse document frequency".

Thus both shared vocabulary of the social network and actors' network position are factored in in the results of the textual analysis of Condor. In the next three sections we will describe our three-step approach: "What – Who – How". "What" stands for the concepts we are extracting and measuring over time. The "Who" represents the actors using the concepts we want to track, while the "How" measures the positive or negative sentiment in which the actors use the concepts. Determining the social network position of actors – the "Who" – has different semantics for each of the three information spheres. On the Web, the betweenness of actors is measured by the linking structure of the Web pages pointing

back to pages talking about them. In the blogosphere, we only consider links to other blog posts as a measure of confidence of other bloggers into the poster of the original blog post. In online forums, the relative importance of a poster is based on the communication structure and the poster's position in the social network.

V. WHAT - MEASURING TEMPORAL BETWEENNESS OF CONCEPTS

The first step to measuring a trend is the tracking of a concept's relative importance in a relevant information sphere – Web, blog, or online forums. 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.

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 networkis a direct indicator of its influence [23]. 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 theWebsites 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 in the chosen information sphere. 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.

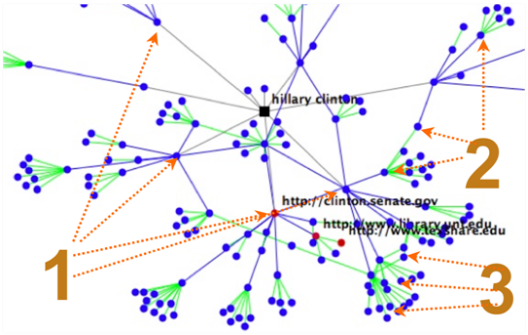

Figure 2. Degree-of-separation search for "Hillary Clinton"

Fig. 2 illustrates the two-mode network map returned to the query "Hillary Clinton". The level-0 node is the query term, level-1 nodes are the URLs connected directly to the query, i.e. the original search results. Level-2 nodes are the most highly ranked search results returned by the "link" query, to each of the top N level-1 nodes. Level-3 nodes are the most highly ranked nodes returned by the "link" queries of each of the level-2 nodes.Fig. 2 gives a visual overview of the betweenness of each of the level-1 and level-2 nodes. The more links a node has pointing to it, the more between it is. For example the node labeled http://clinton.senate.gov is linked by a group of level 2 nodes which themselves are linked by groups of level-3 nodes. This indicates that the node http://clinton.senate.gov will have fairly high betweenness itself.

Fig. 3 illustrates how degree-of-separation search can be used to compare the relative importance of the concepts "gun control", "abortion", "gay marriage", and "Iraq war". This means the importance of an individual concept depends on the linking structure of the temporal network and the betweenness of the other concepts in the network. Condor queries for each concept were run on the Web in 2006, when the war in Iraq was dominating US headlines. Fig.3 shows the semantic social network combining the search results for these four concepts.

A degree-of-separation search for several concepts always results in a fully connected graph since Websites such as Wikipedia or New York Times connect all resources. This is because usually among the level-1 nodes, but at the latest among the level-2 nodes, there will be Wikipedia and other top-rankend Web sites, acting as connectors.

Betweenness values for each concept are calculated in the connected graph formed by combining the Web links pointing to the top ten search results for each of the four Web queries by running a degree-of-separation search for each of the four search queries.

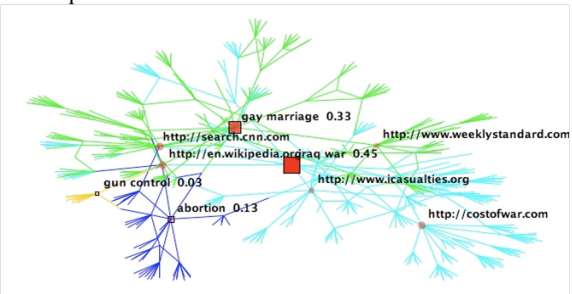

Figure 3. Comparison of the importance on the Web of „gay marriage", "gun control", "abortion", and "Iraq war". Squares are query terms, circles are URLs, size denotes betweenness

The war in Iraq dominates the discussion, followed by gay marriage. Gun control was almost a non-issue at that time,with a centrality factor less than a tenth of the war in Iraq.We can also see that costofwar.com, www.weeklystandard.com, Wikipedia, and cnn.com are the Web sites with the highest betweenness centrality.This also explains why we get a fully connected graph when we combine the four networks for the four concepts: there are always very central, i.e. highly between Web sites such as Wikipedia connecting seemingly unrelated concepts,thus permitting us to calculate betweenness for each concept in comparison to the others.

Note that this ranking has nothing to do with the absolute number of search hits returned by the search engine. If a concept has been around for a long time, it will have accumulated many Web pages, therefore leading to many hits. A newly emerging "hot" concept, which appears on high-ranked Web sites, will not necessarily have that many hits, but will have high betweenness.

Measuring trends is not restricted to measuring popularity of abstract concepts, but can easily be applied to measuring popularity of people. The next example illustrates the "Web popularity" among the top seven Republican and seven Democratic contenders to become the next US President, as of end of August 2006. Fig. 4 shows the combined degree-of-separation search results for 10 US Presidential hopefuls. Each of the colors identifies the set of nodes and links between them retrieved from the information sphere for one of the presidential candidates, e.g. the Web sites and links returned to concept "Al Gore" are shown in blue. While the red squares represent the search queries the red nodes are the Web sites returned by more than one query. The bigger a node the more important it is in the relative network. The relative position of two concepts inside the network to each other can be interpreted as "how close in substance" two concepts, i.e. two presidential candidates are to each other.

Figure 4. Degree-of-separation searches combined for presidential hopefuls in Aug 2006

For example, in fig. 4, Rudolph Giuliani and Newt Gingrich seem to go off together "to the far right". Table I lists the results of the two most recent presidential polls as of end of August 2006 and compares them with the betweenness values of the candidates on the Web calculated in September 2006.

TABLE I. POLLS AND RELATIVE WEB BETWEENNESS FOR US PRESIDENTIAL CANDIDATES IN 2006

Democrats	**Pew Aug 9-13**	**Am.Polling June 13-16**	**Betweenness Web Aug 26**
Hillary Clinton	40%	36%	0.05
Al Gore	18%	-	0.10
John Edwards	11%	15%	0.10
John Kerry	11%	13%	0.05
Joseph Biden	6%	4%	0.02
Bill Richardson	4%	5%	0.06
Russ Feingold	2%	6%	0.01
Republicans			
Rudolph Giuliani	24%	21%	0.09
Condoleezza Rice	21%	30%	0.04
John McCain	20%	20%	0.03
Newt Gingrich	9%	8%	0.05
Mitt Romney	4%	7%	0.02
George Allen	-	5%	0.03
Bill Frist	3%	2%	0.06

Based on the poll values in table I, we would expect Hillary Clinton and Rudy Giuliani to be the most between actors in our Web analysis. The result is slightly different, however. While there are no surprises for Rudy Giuliani, Hillary is not really the top ranked democratic candidate by betweenness. This honor falls to Al Gore and John Edwards, who are tied for first place. The reason for non-candidate Al Gore's surprising popularity were the recent launch of his new movie "An Inconvenient Truth" about global warming, generating buzz for Al Gore not only as a politician, but also as a movie actor and environmentalist. Al Gore therefore connects different Web communities, or in the language of social networks, he bridges structural holes, leading to high

betweenness. Al Gore's high betweenness also illustrates that comparing relative betweenness only makes sense among similar concepts – such as US Presidential candidates in our example.

Repeating the calculations periodically over time permits to measure changes in betweenness of the different candidates to identify trends. This temporal concept importance is the foundation for steps two – "Who" and three – "How" of our approach.

Fig. 5 illustrates the changing betweenness values of the 14 presidential contenders over 14 days. As the blue line shows, non-competing candidate Al Gore's lead is growing, while other leading democratic candidate John Edward's fortunes are declining. The big winner of the first week is candidate Russ Feingold, whose absolute betweenness and thus Web popularity is more than doubling before going down again in the second week. Leading republican candidate Rudy Giuliani is keeping his lead, in a neck-on-neck race with Al Gore. The overall centrality of the combined group analysis is slightly diminishing over the time period, indicating that there is no clear leader emerging thus far.

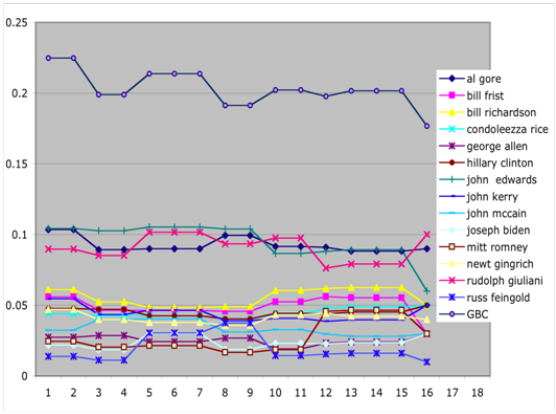

Figure 5. Web buzz trend over 18 days in August 2006 of US Presidential candidates

Fig. 6 illustrates the Blog buzz right after the US presidential elections Nov 4, 2008. Democrat Barack Obama won the elections against Republican John McCain with a landslide in electoral votes (365 against McCain's 162) and 53% of the popular vote. Fig. 6 illustrates this process measuring the betweenness of search strings "John McCain" and "Barack Obama" in the blogosphere. The upper left window shows the minute-by-minute readings, which, at the time of the election, change by the minute based on new posts about either of the candidates on high betweenness blogs such as the Huffingtonpost or Powerlineblog. The overall trend favoring Barack Obama, the blue line, can however clearly be seen. In the accumulated graph, in the upper right window, starting in September, Obama's betweenness line consistently trumps over McCain's betweenness. The bottom left picture shows the social network of blog posts. The blogs talking about McCain form a far more compact cluster, at the very bottom with a tightly interlinked structure. The democratic blogs, linking to Obama, are much wider spread out, and also exhibit fewer interconnecting links, reflecting the wider political interests of the voters supporting Obama. The pie chart at the lower right shows the relative betweenness of the two candidates, 53% for Obama, against 47% for McCain). Note that these relative betweenness numbers correspond to the percentages for the candidates in the popular vote.

VI. WHO - WEIGHING DISCUSSION CONTENT BY THE SOCIAL NETWORK POSITION OF ACTORS

The "Who" step is based on the idea that what certain people say carries more weight, i.e. that some people are more influential than others. As an approximation of their influence we use their betweenness. In the "Who"-step of our approach we add a context-specific weight of the concept's importance, based on the importance of the actor, which is using the phrase. Depending on the information sphere, the actor is either a Web site, a blog (standing in for the respective blogger), and the poster in the online forum. Thereby we factor in that not all actors in the network are equal and that their importance matters for the discussion of the concept.

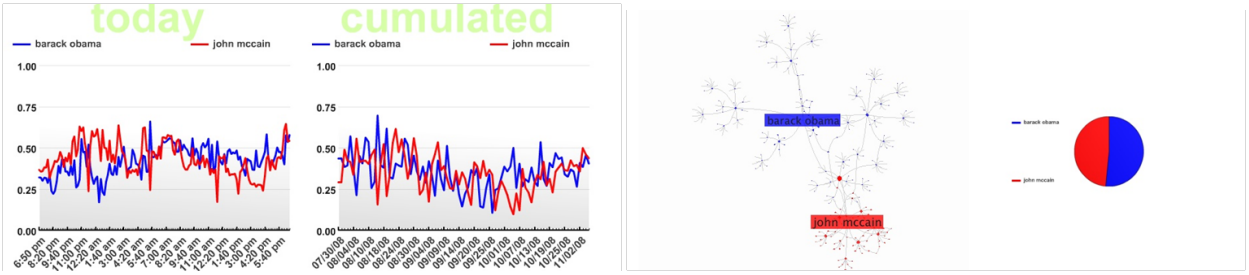

Figure 6. Blog buzz trend with Condor right after November 4, 2008 of US Presidential elections

The context-specific importance of a phrase is based on its originator's betweenness centrality. By multiplying the betweenness centrality of the actor with the betweenness centrality of the concept we factor in the influence of the actor in the information sphere. This not only supports the

elimination of spam that might have been produced to "game the system", but also introduces "expert ranking".

Looking at the presidential candidates example (Fig. 4, 5 & 6) the evaluation of the semantic social network by betweennesspermits to find the most relevant Websites. These Websites can be regarded as "kingmakers" in our context. Kingmakers are Web sites that, through linking to a concept, increase the betweenness of the original concept through their own high betweenness centrality. In our presidential polling analysis, en.wikipedia.org and www.ovaloffice2008.com are the most between Websites. While it is not surprising that Wikipedia is very central, as all candidates take care to get their profiles entered and updated there, the central position of ovaloffice2008 comes as somewhat of a surprise. For each individual network generated by the degree-of-separation search for each candidate, Wikipedia, the candidates' own Websites, and the sites of national newspapers such as the New York Times or the Washington Post rank higher. If the Websites returned to the 14 different degree-of-separation queries are combined, however, a different picture emerges, with Wikipedia and ovaloffice2008 by far having the highest centralities. While the Google page rank of Wikipedia is 9 (out of 10), ovaloffice2008's Google page rank [5] in August 2006 was only 5. Its betweenness in the context of presidential elections, however, is the second highest of all Websites included in this analysis of presidential hopefuls. Ovaloffice2008 also includes a very active forum where citizens of different inclinations and party colors discuss strengths, weaknesses, and chances of the various candidates, motivating the central position of this Website. Note that we will always get one connected network when combining the individual Web networks, as there are the "superconnectors" like Wikipedia and New York Times, linking the individual networks of the candidates.

VII. How - Determine Discussion Quality through Sentiment Analysis

Measuring temporal betweenness of concepts in online communication and weighing the content with the importance of actors provides new possibilities of identifying and analyzing new trends. However, there is a third component that needs to be incorporated into the process. It is not only about What and Who, it is equally important to look at positive and negative emotions in the discussion. For our final example we loaded communication data for 21 stock titles from the finance-related online community Yahoo! Finance into Condor. Yahoo! Finance offers individual message boards for several thousand companies from various industries. The way actors are talking about a particular subject can be determined through sentiment analysis. Besides visualizing and measuring temporal betweenness of concepts or actors Condor includes effective text analysis methods. Through its content process functionality the software automatically can identify most frequent words and word pairs in large amount of texts. [7] has shown that automatic extraction of words and word pairs leads to more precise results than manually selecting positive and negative words. We have implemented a two-step approach first using the automatic term extraction algorithm of Condor to get most relevant words and word pairs. In the second step we created term lists of words and word pairs with positive and negative sentiment by reading the extracted bags of words. The lists are concept dependent and were specifically selected for the analyzed company. Condor provides the possibility of applying stop word lists to exclude common words like "the", "for" or "and". After the identification of positivity and negativity lists we then extracted the frequency of company related, positive, and negative terms within posts. The combination of these three metrics – frequency, positivity, and negativity – represents the sentiment of the forum users on a company. The following table shows the term lists for Goldman Sachs. To enhance the significance and accuracy of the text analysis we implemented an algorithm based on regular expressions. The algorithm detects and analyzes co-occurrence of company terms with positive or negative words in a forum post. This makes it possible to identify the sentiment about a subject in a forum on a particular day.

TABLE II. Company terms, positivity and negativity lists for Goldman Sachs

Company terms	Positivity list	Negativity list
gs, goldmansachs, goldman, sachs	better, bought, buy, buy puts, buy shares, buy stock, buy stocks, buying, earnings, going higher, good, good time, higher prices, investment, long, longs, profits, won	back, bad, didn, dont, down, going down, inflation, little, losses, lower, market down, recession, sell, selling, short, short position, shorting, shorts, sold, stock down

The approach follows the basic "bag-of-words" approach which is also considering co-occurrence of keywords in sentences or text [17]. 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 "Goldman 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. Reference [16] states 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".

VIII. Combining the "What-Who-How": the Web Buzz Index

To test our approach combining social network data from all three information spheres, we collected data over 213 days (April, 1st 2008 until October, 30th 2008) on 21 stock titles on Yahoo! Finance. Additionally, we tracked the temporal Web and blog betweenness for the same titles with Condor. We implemented an algorithm that determines correlation between

the Web buzz and actual stock price. The Web buzz is comprised of Web and blog betweenness, and forum sentiment. Forum sentiment is calculated through the metrics introduced in section 7: term frequency, positivity, and negativity. Each of these metrics has been calculated in two different ways: the simple way only considering sentiment and a second way weighing the sentiment with the social network position of an actor. This makes it possible to weigh forum posts by the "importance" of the poster. This classification results in eight indices, Web betweenness, Blog betweenness, Positivity, Positivity betweenness, Negativity, Negativity betweenness, Wordcount (representing frequency), and Wordcount betweenness.

We calculated index values for a time window of 30 days. We smoothened the index curves by moving averages from five to twelve days. The results for Goldman Sachs are shown in fig 7.

We observed that on days where the stock price rose the negativity indices were inversely correlated. The same was true for the positivity indices on days where the stock price fell. This means that on days with rising stock price the inverse of the negativity index has to be taken, while on days with falling stock price the inverse of the positivity index was taken. As fig. 7 illustrates at time window size of 30 days average correlation values are significant at levels of 0.05 ($n = 30$, $r > 0.361$). We found that the highest value for the moving average most often showed optimal results. Generally, the correlation values of the indices in time window 30 show that a relation between Web buzz and stock price movement exists.

Fig. 8 shows the individual plotted curves cumulatively making up the Web buzz index in relation to the stock price for a moving average of 12 days.

Time Window: 30 days									
	Moving Average								
	5	6	7	8	9	10	11	12	max
Index Type	Average correlation of sub time periods with stock prices								
Wordcount	0.350	0.369*	0.384*	0.392*	0.392*	0.388*	0.388*	0.394*	0.394*
Wordcount Betweenness	0.421*	0.433*	0.440*	0.442*	0.442*	0.441*	0.440*	0.448*	0.448*
Positivity	0.334	0.351	0.361*	0.366*	0.371*	0.374*	0.380*	0.386*	0.386*
Positivity Betweenness	0.409*	0.416*	0.421*	0.424*	0.425*	0.424*	0.424*	0.430*	0.430*
Negativity	0.331	0.344	0.357	0.361*	0.366*	0.368*	0.372*	0.380*	0.380*
Negativity Betweenness	0.406*	0.412*	0.417*	0.417*	0.417*	0.416*	0.415*	0.420*	0.420*
Web Betweenness	0.321	0.322	0.317	0.313	0.314	0.331	0.342*	0.350	0.350
Blog Betweenness	0.348	0.370*	0.383*	0.394*	0.403*	0.414*	0.424*	0.427*	0.427*
(α=0.05)*									

Figure 7. Correlations between stock price and the different components of the Web Buzz Index for Goldman Sachs

IX. Outlook and Conclusion

In this paper we have shown that buzz on the Web mirrors the real world. Tracking concepts on the Web by differentiating between the Web at large, blogs, and online forums, and combining what people say with their social network position indeed permits to discover trends, frequently before the real world has become aware of them.

There remain some issues that deserve further investigation. The first concerns our dependence on the rankings of the search results by the search engine. We have used Google, Google Blog Search, MSN Search, and Yahoo.

While the top n Web sites about a topic returned by the different search engines vary, we found surprising consistency in the relative betweeness values of the search topics. We explain this through the presence of central Web sites such as Wikipedia, Yahoo, and the New York Times Web site in the resulting link networks. These Web sites always come up in the searches at one or two degrees of separation to the search topic, providing a consistent linking structure. The second issue is about causality. While we have demonstrated correlation between Web buzz and real-world events and have demonstrated the predictive capabilities of our approach for political elections and Oscars [19], more work needs to be done to formally show causality for stocks.

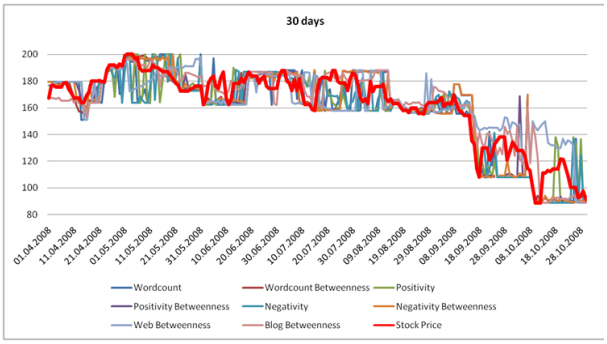

Figure 8. The 8 Web Buzz Indices plotted against the stock price of Goldman Sachs

We are currently testing our system in different application areas, trying to increase the accuracy of our political predictions and stock trend correlations. Possible extensions of our approach are the addition of the concept of fading in and out of new ideas. Frequently, new ideas are brought up by

visionary people, only to lay dormant for extended periods of time until they are finally picked up by larger groups of people. We speculate that extending our model to incorporate this process might increase the correlation between Web buzz and the real world events we are trying to track. We also intend to explore whether different groups of actors based on their network position have different influence on Web communication. This would ultimately also lead to more accurate predictions. We are currently improving our sentiment extraction methods with additional algorithms, e.g. dynamically enhancing positivity and negativity lists with machine learning techniques. To further increase the predictive quality, we consider approaches such as applying a dynamic time offset between the Web buzz index and the stock price, and including industry indices and trading volumes.

Another idea is to combine the Web buzz analysis with prediction markets [25], by setting up automated agents trading in prediction markets based on Web buzz analysis. Extending this line of research, human participants in prediction markets could be given access to our Web trend prediction results in order to increase the quality of the prediction market.

Our vision is to develop a general system for trend prediction, identifying new ideas early on while they are being raised by the trendsetters. At this stage, new ideas have not yet been recognized by the rest of the world, but discovering them can be extremely valuable. Applications of our system might be for politicians trying to find out what the real concerns of their constituency are, or for financial regulators trying to identify micro- and macro-trends in financial markets.

ACKNOWLEDGMENT

We would like to thank Hauke Führes for patiently adding all our change wishes to Condor for the Web Buzz Index. We are indebted to Manfred Vogel for providing us with an excellent IT infrastructure at University of Applied Sciences North West Switzerland in Brugg, and to Tom Malone for insightful discussions on the properties of gatekeeper words in semantic networks.

REFERENCES

[1] Adar, E., Zhang, L., Adamic, L., Lukose, R. (2004) Implicit Structure and the Dynamics of Blogspace: Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference.

[2] Al Hasan, M., Chaoji, V., Salem, S., & Mohammed Z. (2006) Link Prediction using Supervised Learning: Proc 2006 Workshop on Link Analysis, Counterterrorism and Security.

[3] Antweiler, W., Frank, M. (2004) "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", The Journal of Finance. Vol. LIX, No. 3.

[4] Barabasi, L. (2003) Linked: How Everything Is Connected to Everything Else and What It Means. Plume.

[5] 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.

[6] Chakrabarti, S., Joshi, M., Kunal, P., Pennok, D. (2002) The Structure of Broad Topics on the Web.Proc, WWW 2002, Hawaii.

[7] 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.

[8] De Choudhury, M., Sundaram, H., John, A., Seligmann, D. (2008) Can Blog Communication Dynamics be correlated with Stock Market Activity? Hypertext 2008, Pittsburg: PA – 19 to 21

[9] Dodge, M., Kitchin, R. (2002) Atlas of Cyberspace, Pearson Education.

[10] Dodge, M. Kitchin, R. (2000) Mapping Cyberspace, Routledge.

[11] Fruchterman, T. M. J., Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11).

[12] 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)

[13] 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).

[14] Jones, A. L. (2006) Have internet message boards changed market behavior?, info, Vol. 8 No. 5 2006, pp. 67-76.

[15] Kleinberg, J. (2008) The Convergence of Social and Technological Networks, Communications of the ACM, Vol. 51 No. 11 November 2008, pp. 66-72.

[16] 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.

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

[18] Richardson, M., Domingos, P. (2002) Mining Knowledge Sharing Sites for Viral Marketing, Proc. ACM SIGKDD, 2002.

[19] Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (5): 513–523.

[20] Tapscott, D.; Williams, A. D. (2006), Wikinomics: How Mass Collaboration Changes Everything, Portfolio Hardcover, New York, 2006.

[21] Tetlock, P. (2007) Giving content to investor sentiment: The role of media in the stock market, Journal of Finance, Forthcoming.

[22] Tumarkin, R., Whitelaw, R. F. (2001) "News or Noise?": Internet Message Board Activity on Stock Prices, Financial Analysts Journal, 57: pp. 41-51.

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

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

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