Decision Support Systems xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Blog mining-review and extensions: "From each according to his opinion"

Daniel E. O'Leary

Marshall School of Business, University of Southern California, Los Angeles, CA 90089-0441, United States

ARTICLE INFO

Available online xxxx

Keywords:
Blogs
Blog mining
Financial blogs
Sentiment
Corporate image
Public image
Blogs and Sales

ABSTRACT

Blogs provide a type of website that contains information and personal opinions of the individual authors. The purpose of this paper is to review some of the literature aimed at gathering opinion, sentiment and information from blogs. This paper also extends the previous literature in a number of directions, extending the use of knowledge from tags on blogs, finding the need for domain specific terms to capture a richer understanding of mood of a blog and finding a relationship between information in message boards and blogs. The relationship between blog chatter and sales, and blogs and public image are also examined.

1. Introduction

In the early days of the Internet, most of the information content was generated by companies, governments and universities, however, now individuals generate more than fifty percent of the Internet content [10]. As an example of that individual generated content, as of May 4, 2010, according to blogpulse.com there were 126,861,574 blogs. In contrast to Wikis (encyclopedia-like sources of information generally gathered from experts in a particular area), blogs have evolved as a media that allows the user to present a range of information including personal viewpoints and opinions. As blog information has become available, enterprises increasingly have seen those blogs potentially as an important source of information and knowledge.

With over 126 million blogs, the Internet provides a setting where different individuals provide different information, insights and opinions. As noted by Hayek [12] "…knowledge (is) not given to anyone in its totality." Instead "…the knowledge of the circumstances of which we must make use never exists in concentrated or integrated form, but solely as the dispersed bit of incomplete and frequently contradictory knowledge which all the separate individuals possess." Since information and knowledge distributed in blogs potentially offer firms access to the different perspectives and different insights possessed by numerous individuals, if firms are interested in gathering information about what individuals "think" they need to examine blogs, both individually and in the aggregate since blogs provide a wide range of individual information and knowledge sources.

Importantly, Hayek [12] also distinguishes between "scientific knowledge," defined as knowledge of facts, and "unscientific knowledge," defined as "…the knowledge of the particular circumstances of time and place…special knowledge of circumstances of the fleeting moment, not known to others." Hayek [12] notes that "information" or scientific knowledge, is central to neoclassical economics, where agents often are assumed to possess perfect and identical information. But, Hayek argues, that focusing solely on information greatly oversimplifies the task of explaining economic behavior because it ignores the central importance of unscientific knowledge. Hayek's dichotomy between scientific and unscientific knowledge is similar to Polanyi's [28] distinction between "explicit" and "tacit" knowledge. Explicit knowledge is defined as knowledge that is or can be documented and easily communicated and interpreted. In contrast, tacit knowledge derives from experience and involvement in a specific context, and often only resides "in the heads" of individuals. Tacit knowledge includes individuals' beliefs, mental models, and viewpoints, and thus is inherently difficult to communicate [27]. Blogs allow users to comment on events and other materials potentially facilitating communication of tacit knowledge. For example, we will find that analysis of blogs can provide insight into the "sentiment" or "opinion" (positive or negative) of the blog writer regarding the issue being analyzed. Accordingly, blogs are likely to provide both scientific and unscientific and explicit and tacit knowledge in an explicit format. As a result, blogs can provide an important knowledge management tool.

1.1. Purpose of this paper

Since there are over 126 million blogs currently available, with knowledge dispersed broadly to many sources, and with potentially so much scientific and unscientific, and explicit and tacit knowledge within them, it is important to ask *"What can we learn from blogs?" "How can we determine the 'sentiment,' (positive or negative) about the issue being expressed in the blog?" "Can those blogs, when analyzed*

E-mail address: oleary@usc.edu.

0167-9236/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2011.01.016

individually or in the aggregate, provide insight into attitudes about products, financial information and a range of other entities or events?"

This paper broadly investigates those questions, focusing on how we might capture information and knowledge from blogs in search of key insights. Accordingly, the purpose of this paper is to explore blog mining, both in general and in the context of some specific applications. In so doing, this paper also provides a survey of the literature on capturing sentiment from blogs, mining blogs and how blog information relates to knowledge from other knowledge source sources, such as message boards. In addition, this paper extends the existing literature in five new ways. First, researchers have used blogger supplied mood tags to gage the mood of the Internet. But, researchers also have noted that blogger supplied tags have been decreasing over time, threatening the ability to capture mood. However, this paper notes that an alternative source, reader supplied tags (e.g. DElicious.com) can be used as a basis to replace the blogger supplied tags. Second, the paper analyzes the extent to which one approach (mood words) could be used to capture whether a blog is commenting positively or negatively. I find that one approach to determining the overall mood is to look for statements of disclosure as to mood within the blog. For example, some bloggers indicate that they have a "positive opinion" about some event or entity. However, I also find that a general dictionary does not fully capture domain specific mood information in a financial domain. Third, this paper investigates the relationship between information in two different knowledge sources, blogs and message boards. I find that in a case study using financial accounting information, the information between these two sources is highly correlated. Fourth, researchers have found a relationship between blog "chatter" (activity) about products (movies, books, music, etc.) and the sales of those same products. As a result of those findings, the implication is that firms need to create blog activity about their products. However, this paper questions the causality link and whether firms will be able to generate sales if they generate chatter. Fifth, this paper analyzes why and how firms potentially monitor their "public image" through mining blogs. I find a number of limitations of using a general concept of "public image" and suggest that specific components, e.g., "going green" be the primary focus of such analyzes.

1.2. This paper

This paper proceeds in the following manner. Section 2 provides a brief background on blogs, including blog search, blog mining, and noting the impact of the importance of blogs. Section 3 investigates different potential samples of blogs that might be analyzed. Section 4 analyzes how we might gather data about opinion from blogs, including using frequency of appearance and tags describing content. Section5 examines basic research into finding sentiment and opinion in blogs and provides a test of the use of opinion words as a means of finding opinion in discussions about stocks in financial applications. Section 6 examines the determination of information from message boards, a distinct, but apparently somewhat similar knowledge source, and tests the similarity of the information derived from one message board study to the information in blogs and financial blogs. Section 7 investigates some of the literature that suggests that blog chatter and sales are related; and that section also discusses some extensions to that literature. Section 8 analyzes using blogs to gather information about a firm's public image, and summarizes some extensions to that literature. Finally, Section 9 briefly summarizes the paper and investigates some extensions.

2. Background: blogs, blog search and blog mining

Blogs are websites that provide content often generated by individuals. An analysis of "BlogPulse.com" provides many examples of the types of topics that are found in blogs: financial, political, entertainment, and news. Virtually, anyone can blog. There are few filters in place to limit blogs or what is in them. As a result, there are millions of blogs. Accordingly, substantial information is put in the so-called "blogosphere," of which much may be redundant and correlated. However, since the information is coming from so many different sources, at so many different times, there may be real information, not previously realized or recognized, that is embedded in the blogs.

Blogs are likely to represent a single individual or a group. Blog information may differ from other kinds of text. For example, blogs are not likely to be as well edited, as newspaper or magazine text. In contrast to other forms of text, blogs may use incomplete sentences and phrases.

Individual blogs can have a huge impact and bloggers can gain substantial notoriety. For example, the Korean Blogger Dae-sung Park was widely read under the pseudo name "Minerva." News accounts apparently suggest that Mr. Park's blog postings had led to a plunge in the value of the Korean Won, forcing the government to intervene in trading [29]. As a result, South Korean officials arrested Park and shut down his blog.

Blogs have generally been associated with the advent of what has been called Web 2.0, emerging after the first wave of web innovation. In addition, increasing focus is being placed on capturing semantic knowledge about the blog, interactive sharing of information and the corresponding collaboration that such sharing can bring.

2.1. Blog search

A number of search engines, including Technorati (http://technorati.com/blogs/directory/) and Google, provide the ability to search blogs for specific concepts or issues. These search engines allow users to easily find blogs that contain pre-specified chunks of opinionated text. For example, "X SUCKS" would allow finding all of the pages with the appropriate set of opinion-oriented text. Such search engines can be employed by other software to generate information and insights.

2.2. Blog mining

Blog mining is the process of searching and analyzing blogs in order to generate additional insights that might otherwise not be found by examining a single blog. If blogs contain information and knowledge, whether tacit or explicit, by analyzing and "mining" the information in them, we can begin to make assertions, particularly in those settings where we are able to pull together information and knowledge from multiple different blogs. Blog mining tries to create an overall understanding of information from the disparate sources.

Marketing researchers and companies have long been interested in capturing information and knowledge about the opinions of buyers or potential buyers of their products. However, interviewing people about their opinions is time consuming and costly, and there is concern if the individual is telling the truth or telling the marketer what they want to hear. In contrast, blogs provide a readily available and opinion-based content media that provides sentiment about a range of issues. Further, that qualitative content can be matched against key performance indicators, such as sales, profits or stock price. As a result, being able to use those blogs for gathering opinion information potentially can provide a low cost source of information about those opinions and sentiment, regarding particular issues and concerns, gathered in real time.

2.3. Blogs and organizations

In many cases, organizations may have structures that function as "sensors" in the environment to capture feedback from clients (e.g., customer service and customer relations). However, in some cases those organizational devises do not work. For example, in the recent case of Toyota, it was suggested that information did not always work all the way up the organizational hierarchy but instead was "stuck" at

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

different places [8]. As a result, an alternative is to have readily available access to opinion information at any point in the hierarchy. Being able to gather information and knowledge from blogs could facilitate broader organizational access to outside information. Ultimately, this could result in changes in organizational structures, e.g., flatter organizations, because the new flow of information may limit the need for existing hierarchy.

Further, organizational sensors may be limited by type of information that they capture. For example, customer service generally can capture information about what is wrong with an existing product, but not what could be added to existing products to make them better or help them mitigate existing problems. As a result, being able to gather information from blogs about how products might be changed could facilitate information flows and broaden the base of information access.

In addition, in today's world, because of the Internet, information can explode from user to user in real time. Within hours of an email or a blog being posted, information can be diffused to millions of people. For example, Vara [34] discussed how information about using a ballpoint pen to open a Kryptonite bike lock spread around the world. Within two days information was in blogs, but it was not until a few more days that the company responded. However, by then, there were angry and confused customers flooding the company with questions.

> *"That was probably the most astounding thing — to see how rapidly this whole thing developed and moved around the world at an amazing speed," says Karen Rizzo, director of marketing at Kryptonite, a division of Ingersoll-Rand Co. Ltd.* [34]

If companies are aware of trends of opinion or sentiment they can act to leverage opportunities or respond to problems in a timely manner. Since blogs are constantly being published, searching and analyzing those blogs can provide timely information to potentially head off problems.

Further, insight into sentiment associated with blogs can provide potential investment opportunities. For example, perhaps information in blogs can provide insights into the investment of stocks or commodities. Frequently, bloggers provide comments and insights into enterprises, and those comments and insights may influence stock prices.

2.4. Blogs and competitors

Blogs can serve as an important part of competitive analysis. In particular, blogs can provide companies with insights into their competitors; in terms of what is going well and what is not going well, thus providing insight into potential market opportunities. Competitor blogs and blogs about competitors can also be used to anticipate concern that others may have with the firm. For example, if a competitor has a quality of product concern it probably will not be long before the quality of the particular firm will come under scrutiny, along with others in the same industry.

3. Which blogs are analyzed?

A critical decision facing a company or researcher that is doing blog mining is "Which blogs are to be analyzed?" The literature seems to largely ignore this issue, but the choice should be driven by the objective and available resources. There are a number of different choices: a small selected set of blogs, a random set of blogs, all available blogs, blogs of a particular type, blogs from a particular time period, or an experimental set of blogs.

3.1. Selected sample

One approach is to choose a selected sample of blogs. For example, if a firm is concerned with the opinions of some specific people or groups then it can be beneficial to focus on that specific subsample.

> *BuzzMetrics has created a panel of what it calls "word-of-mouth influencers," a list of thousands of bloggers, message-board posters and other people BuzzMetrics has deemed influential in the online community, in part by examining traffic numbers. By studying their online interaction, BuzzMetrics says it can give companies important information about how they are perceived by customers.* [34]

There are some advantages of this approach, including focusing on a particular sample of blogs that are known or at least thought to be important to the set of information being analyzed. In addition, as long as it is a relatively small sample, this approach will not require a substantial time for analysis. The primary disadvantage of this approach is that there may be important information in those blogs not analyzed.

3.2. Random sample

Another approach is to choose a random sample of blogs to analyze. Using this approach can facilitate getting a broad base of alternative opinions, while still making the number of blogs analyzed computationally feasible. Further, statistical methods might be employed to determine how large a sample would be necessary to draw statistically significant (at whatever level specified) conclusions about the data.

3.3. All available blogs

Search tools, such as Google Scholar or Technorati, can identify all or almost all of the blogs that meet certain criteria specified for an analysis. Such searches enable one to obtain a very large (almost complete) set of opinions contained in blogs at some specific point in time. However, the key disadvantage is the need to review or analyze thousands, hundreds of thousands, or even millions of blogs.

3.4. All available blogs of a particular type

At Blogsearch.Google.com page blogs are categorized according to a number of content categories, including Politics, World, US, Business, and Technology. One approach would be to focus on those blogs in a particular function, expecting a certain amount of homogeneity among the blogs. For example, in an experiment discussed below, an analysis of stock market language likely could focus on the financial blogs.

3.5. Time-based selection

Another issue is the role of time. In the case of the analysis of financial information, interest may only be in the most recent information about the firm or in the information over the last 12 months. Alternatively, in the case of an archival study, data across a longer time period may be preferred.

3.6. An experimental set of blogs

In some cases, the set of blogs of concern is a specific test collection. For example, the 2006 TREC Blog track, organized by NIST, asked participants to implement and evaluate a system to do "opinion retrieval" from blog posts. Specifically, the task was defined as follows: build a system that will take a query string describing a topic,

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

e.g., "March of the Penguins", and return a ranked list of blog posts that express an opinion, positive or negative, about the topic. Systems would be ranked using a number of evaluative measures, including recall and precision for their ability to correctly identify opinion (http://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf).

3.7. Source of blogs: internal vs. external

Up to this point in time, the focus has largely been on blogs generated externally to the firm. However, in the case of an enterprise, blogs can be either external to the company (e.g., from customers) or internal (e.g., internally generated by employees). Internally generated blogs can also be a source of information for the company. Further, in some cases, systems that ultimately focus on external blogs originally were developed for analysis of internal text systems [14].

4. Opinions and sentiment

Humans generally are able to distinguish between positive and negative opinions, although the case of sarcasm can make it difficult. However, it is difficult for humans to distinguish between neutral positions and opinion bearing positions [15]. As a result, the goal for a computer program is to determine a similar set of perspectives, but it is more likely that such programs will be able to determine positive or negative positions.

4.1. How do people determine opinion?

Toulmin [32] set forth the following approach in his work, "…we shall be studying the operation of arguments sentence by sentence, in order to see how their validity or invalidity is connected with manner of laying them out…" However, such an approach may be very difficult to capture in a computer program. As a result, alternative approaches have been employed ranging from frequency of appearance, to appearance of particular words, to phrases and sentences.

4.2. Frequency of appearance

Frequency of appearance of terms is widely used as a measure of interest in a topic in blogs (e.g., BlogPulse.com and others). But is frequent appearance good or bad? One perspective is that any publicity is good publicity, but there are different opinions as to whether or not that is true. (http://www.answers.com/topic/any-publicity-is-good-publicity; also consider Rhonda Farr's classic statement, 'Publicity, darling. Just publicity. Any kind is better than none at all.' [1933 R. Chandler in *Black Mask* Dec. 26] and the Washington Times article [2002, May 9th, page C8] on how Mike Tyson may have disproved the adage that any publicity is good publicity).

One approach to the analysis of blogs or for that matter any written word is to analyze the frequency of appearance or incremental frequency of appearance of particular words or phrases. Asur and Huberman [2] recently found that simply number of appearances of discussion about a single topic can be used to predict characteristics of the topics. As an example, I did a search on "Microsoft" and "IBM" using Google's search engine for blogs and got the following number of pages associated with each company, given in Table 1.

Based purely on the number of occurrences, there appears to be substantial opinion and information being expressed about those two companies. Further, for both firms, the number of pages substantially increased in little more than a month, also suggesting incremental information. Finally, Microsoft apparently had a higher percentage increase in its appearances, as compared to IBM suggesting greater opinion and information being expressed about Microsoft, at that time.

Frequency has some important structural bases that need to be taken into account. For example, as seen in Fig. 1, both IBM and Microsoft appear to exhibit day of the week effects, specifically weekday as compared to week-end differences in the frequency of terms being mentioned.

Table 1
Number of blogs mentioning Microsoft and IBM.

	January 30, 2010	March 9, 2010	% Increase
Microsoft	76,393,602	96,907,544	26.853
IBM	15,472,510	16,870,997	9.038

However, greater frequency of appearance does not necessarily mean more information about an entity or event. For example, there can be multiple posts of the same page or there can be spam posts that contain no information. Spam blogs, also referred to as "splogs" are nonsensical blogs with multiple search terms embedded in them [24]. Splogs are designed to induce readers to click ads and to get search engines to capture the number of links. It has been estimated that more than 50% of the blogs are splogs [24]. Accordingly, the issue of being able to determine if a blog is a splog is increasingly important.

4.3. Tags on blogs

From this discussion we can note that simple appearance numbers of concepts in blog texts can suggest that there is information about a concept and that the relative, incremental increase can also provide additional insight. However the number of appearances doesn't provide information on sentiment or on whether an opinion expressed is positive, negative or neutral. In order to assess whether a blog is expressing more positive or more negative opinion, we can either examine tags placed on the blogs or directly examine the content to try and determine whether the blog is positive or negative.

One approach to generating opinion information about a blog is to have the blogger indicate the mood, sentiment or opinion of the blog. For example, a large number of "LiveJournal" bloggers have tagged their posts with mood information, using any of 132 predetermined moods. As a result, some researchers have used that notion to try to get a reading on "moods in the blogosphere." This has allowed researchers to track the mood of bloggers over time, providing a time series of those blog moods [4]. Those researchers have found a number of interesting results. For example, weather and holidays influence blog moods and several moods appear to be cyclical or seasonal.

However, blogger supplied mood tags are not the only tags that can be analyzed. Analysis of tags can be extended to other tag information, such as tags from "DElicious" and other sources. Examining tags leads to many potential words suggestive of moods, positive, negative and other. For example, a search on "stupid" also led to other tags, including "funny," "humor," "stupid," "spam," and some others. As a result, we do not need to depend on bloggers to label their own blogs, but can use labels assigned by others. This is particularly critical since there has been a definite decrease in blogger supplied labels over time, but interest in DElicious and other tags continues to grow [4].

Further analysis found that 78% of the first 100 items that were listed on DElicious under the term "stupid" also included "fun" or "funny" or both. This has a number of implications. First, using multiple tags, we can get gradations associated with the moods, or mood concepts, e.g., "stupid" and "funny" vs. "stupid" and "sad." Second, this also suggests that labels can be correlated. Empirically, "stupid" and "funny" co-occur in DElicious. Future work could further investigate what words co-occur on human generated tags and how such tags can provide detailed understanding of moods.

4.4. Additional factors: blogger and site consistency

A different approach to determining blog mood is to investigate the mood information using additional factors related to a blog, for

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

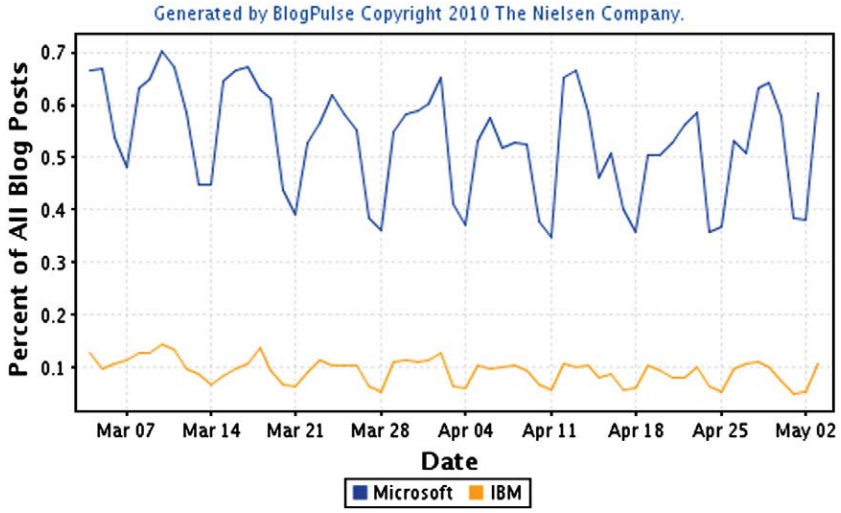

Fig. 1. Characteristics of blog postings.

example blog authors or blog sites or blog site names. This approach would use the notion that some bloggers are relatively consistently negative, positive or neutral in their blogs. Further, that consistency could extend to the site level, as seen by the many sites with labels such as "www.hateXXX.com," "www.i-hate-XXX.com" or "www.XXXdirtylaundry.com." Finally, this would recognize that the name of the blog site could provide additional insight. Accordingly, additional information, other than the content, also can be used to gather information about a blog's mood, sentiment or opinion.

5. Examination of content: sentiment and opinion in blogs and text

One of the key tasks is the very definition of what makes something opinion. Other tasks include determining how opinion is captured and what is the "unit of reference" of the opinion? Researchers have used multiple approaches and a number of different issues have been identified.

5.1. How is opinion captured in blogs?

Blogs provide a forum for bloggers to provide opinion, sentiment and information about a range of issues. Although bloggers may use pictures, Dilbert cartoons, and videos, in this paper we are primarily concerned with the expression of opinion using text.

5.2. What is the unit of reference for the opinion?

In the context of written word, opinion determination can occur at many different levels. For example, opinion can be expressed at the individual word level, the individual sentence level, the paragraph level, a single date blog, or across multiple dated blogs. As a result, since blogs cover multiple bloggers over multiple time periods, there can be multiple "units of reference." Accordingly, when determining the overall opinion or sentiment of a blog or information content, a decision must be made to determine if a word, paragraph, sentence or whatever, determines the opinion or sentiment of the blog.

5.3. Mood declaration: how can we tell if it is positive, negative or neutral?

In the name of clarity, some bloggers may declare whether the particular blog is positive or negative or neutral. On March 9, 2010, I did an analysis to try to begin to understand the extent to which blogs made such declarations, representing positive, negative or neutral opinions. I did a search to identify the number of occurrences of "Neutral Opinion" (3686), "Negative Opinion" (30,336) and "Positive Opinion" (19,515), based on the number of pages identified in a Google Search of blogs with those terms. Further, the search for a joint occurrence of "Negative Opinion" and "Positive Opinion" identified only (503) pages. These results confirm the notion that a blog likely will take a positive or negative approach when generating an opinion. This suggests that when mining for opinion, in addition to other approaches we can effectively ask the blogger, by looking for mood declarations, what they thought was the mood of the blog. Similarly, an analysis of the declarative expressions "looks good" (2,860,168) and "looks good to me" (353,382) apparently are used to declare a positive opinion. Further, the terms "thumbs up" (2,273,209) and "thumbs down" (668,639) also apparently were used to declare blog opinion. Finally, this approach can be used with other approaches to facilitate and confirm the results. For example, we can identify the "declaration" and then determine if we can "substantiate" the opinion based on opinion word occurrences, as discussed in the next section.

5.4. Opinion word dictionaries

Another approach to identify opinions in text is to identify specific words that suggest a particular opinion. For example, the term "X SUCKS" is (very) likely to connote dissatisfaction with "X". In this case, the single word "SUCKS" is an opinionated word.

To what extent can we capture "opinion" and "sentiment" through words? Recently, Magistrates from courts across Italy walked out as a protest against the Prime Minister. As noted by Gioacchino Natoli from the Italian National Magistrates Association Union [17], "An execution squad, sewer, cancer, metastasis — these are some of the words that the prime minister and his deputies have used to describe us." In this case, a set of words, each with a negative connotation was used to describe the magistrates, suggesting that one approach to searching for opinion and sentiment in language has been through generating so-called "opinion word dictionaries" [13,22,34,35]. A sample of an opinion word dictionary is summarized in Table 2. These dictionaries work on the notion that if the written text uses words from the dictionary, then the text's opinion can be categorized according to the word's categorization in the dictionary. Opinion words are not limited to a particular type of word. Opinion words can be adjectives, adverbs, nouns or verbs. The Italian Magistrates described words were nouns. The words in Table 2 are adjectives and verbs.

Implementation of this approach generally looks for opinionated words in proximity of the particular entity or event [3]. For example,

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

Table 2
Sample opinion word dictionary.
Source: Ref. [36].

Positive verbs	Negative verbs	Positive adjectives	Negative adjectives
Love, like	Hate, dislike	Good, best, better, happy, extraordinary, successful, glad, desirable, worthy, remarkable, funny, lovely, entertaining, decent, beautiful, fascinating, brilliant, gorgeous, perfect, nice, fantastic, impressive, amazing, splendid, distinctive, desirable, excellent, great, awesome, and fabulous	Bad, awful, suck, worse, worst, poor, annoying, and stupid

Attardi and Simi [3] did a search on opinion words within 6 words of "George Bush." If the concern was with tracking information about a certain stock, this approach could be implemented as finding opinion words within a certain number of words from that stock name. If the concern was with tracking movies, then the target would be opinion words within a certain number of words from the movie title or movie director or top star.

Opinion word dictionaries have both advantages and disadvantages. Opinion word dictionaries are a straightforward and simple approach to finding what appear to be opinion "markers." As we saw above with the Italian Magistrates, words capture opinion and sentiment and can be used to denote or mark text or conversations as expressing either positive or negative opinion. Further, the quality of the opinion word dictionary can have a large impact on the ability to determine the opinion in a blog. For example [3], "We should note that the so called opinionated words being extracted from a general dictionary are not very specific and include terms such as 'like,' 'hate,' 'not,' but also 'want' and 'wish'."

On the other hand, although some words can appear to be signals as to opinions, in some cases their usage may not be the same as in the dictionary. For example, consider the sentence "X is awfully good." The sentence contains the term "awful" and the term "good." Thus, in this case, an opinion word dictionary could lead to conclude that the same sentence is both positive and negative sentiment. As another example, I found the phrase, "I love Microsoft bashing" in a blog. In this case, there is a positive word (love) and a negative word (bash) around the entity (Microsoft). As a result, it is questionable that words literally are always positive indicators or negative indicators. Further, combinations of positive and negative words around an entity are not always positive or negative.

5.5. A test of opinionated words in finance

Much research on opinion in blogs has used generic approaches, and has not focused on characteristics of specific domains (e.g., [3,4,10,25,26,31]). As a result, it is not clear if the domain (e.g., "finance") would have an impact on the ability of opinion-like words to spot opinions. Further, it is not clear if domains have "special" words that indicate opinion and how those words might be discovered or discerned to help with the search.

In a discussion of semantic search, Hampp and Lang [11] suggested that search ultimately can be tailored to specific industries, such as manufacturing, services or government, in order to fully leverage the ability to spot particular events, agents, times and locations and understand the sentiment associated with those words. They suggest that in these settings, domain specific knowledge can be leveraged to facilitate disambiguation of terms and better understand context. Accordingly, I would hypothesize if "generic opinionated word" search does not take into account domain words, then the search will miss some occurrences related to events in the specific domain.

In order to test the importance of domain in the use of opinionated words, a fund manager was interviewed and phrases that were likely to be used to describe a stock in a negative manner were captured. The resulting list is not comprehensive. However thirty phrases were generated providing a number of opportunities to see if the phrases and blogs that use these phrases would be found if we were to use either of the above two opinion word approaches. The list is summarized in Table 3, along with the number of times that the phrase was found in a Google search of blogs. Analysis of the list suggests that terms like "bearish," "plunged," "dropped like a rock" and some others probably would not have been found if generic opinion words were used. This indicates that some domain-specific words could be used to supplement the generic opinion words suggested above.

6. Related knowledge source applications: Internet stock message boards

A number of researchers have investigated a closely related set of information from an alternative knowledge source, that of Internet stock message boards. Stock message boards are not blogs, but instead a medium that allows users to exchange information about a particular company. Message boards appear to be used extensively in financial applications. Tumarkin and Whitelaw [33] investigated messages on RagingBull.com, focusing on the Internet Service Sector. They found that abnormal positive returns preceded the days where there were strong positive reviews on the message boards. Antweiler and Frank [1] investigated Dow Jones companies associated with RagingBull.com and Yahoo finance and found that message posting helps predict volatility and that higher negative postings helps predict negative subsequent returns. Sabherwal et al. [30] studied messages on "TheLion.com" and found that there were abnormal returns if a stock was one of the 10 most talked about. Lerman [21] investigated Yahoo stock message boards, analyzing the extent to which accounting information was embedded in the messages and she

Table 3
Some financial phrases indicating bad news about a company stock and the number of their occurrences.

"Worst" "stock"	5094464
"Awful" "stock"	361460
"Stock sucks"	335,656
"Stock" "worthless"	185,157
"Seize the company"	164,510
"Stock price dropped"	22,873
"Stock" "going bankrupt"	16951
"Very bearish"	13009
"Stock" "not worth anything"	10326
"Stock price dropped like a rock"	8256
"Stock price" "plunged"	8174
"Stock is a dog"	7447
"Poor stock performance"	5798
"Stock is no good"	5577
"Worst stock"	5064
"Stock" "learned my lesson"	4821
"Stock" "outlook is bleak"	3651
"Huge losses" "stock markets"	3312
"Bearish on this stock"	2260
"Stock price" "huge losses"	1445
"Stock price" "huge loss"	532
"Awful stock"	353
"Bearish convergence"	177
"Stock" "lost my shirt"	113
"Overvalued stock"	69
"Falling wedge support"	10
"Stock is a POS"	6
"Why did I ever buy this stock"	2
"Small cap stocks suck"	1
"Shorters are converging"	1

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

found that accounting related chatter increased around the time of accounting information releases. She also found that discussions increased during times of uncertainty, arguing that this leads to better informed investors.

6.1. A test of the relationship between information in blogs and message boards

Although there has been substantial attention given to analysis of message boards and chat rooms, there has been limited, if any, analysis of the existence of financial information in blogs. Accordingly, our concern is to investigate how we might analyze blogs to determine the existence of relevant information in blogs.

It is widely accepted that message boards are generally seen as a different knowledge source than blogs. There are a number of differences between the two [20]. Perhaps the primary difference is that blogs are controlled by a single group or individual, while message boards reflect the input of a number of individuals. In addition, there generally is greater depth on a blog, as topics are more focused and are not driven by the group. However, the "comments" to a blog may be seen as somewhat similar to those on a message board, since virtually anyone can add comments and those comments are not driven by the blog writer, although they generally are based on the blog.

We will say that there is "knowledge source clustering" when two different knowledge sources provide highly correlated information. One approach to provide an initial test of the similarity of the information content of blogs and message boards is to use a case study and determine the extent to which information generated by the two different media are correlated. Although just a case study, the terms analyzed are very specific to accounting and finance, and can provide us with some insights into the extent to which blogs and message boards are related.

Lerman [21] provides an analysis of the occurrence of accounting words as part a detailed study that mined information from financial message boards. In particular, she investigated occurrence of accounting terms on message boards for both S&P 1500 firms and non-S&P firms, noting, "financial message boards provide a unique medium to analyze investors' attention to accounting information…" To test the relationship between blog information and message board information this paper compares the occurrence of accounting information in blogs to those results on message boards. Since blogs and message boards are seen as two different knowledge sources, a priori, we would hypothesize that the occurrence of accounting terms on blogs and message boards are not correlated.

In order to test this relationship, information regarding the accounting words was gathered from both Google Blogs and Technorati (Finance). Technorati offers not only information about the blogs, but also information about comments on the blogs. Since comments are driven by the individuals responding to the blogs, there may be a closer relationship between the comments and the message boards. The results are summarized in Table 4, and the corresponding correlation coefficients between the two sets of information are summarized in Table 5. The results indicate each of the correlation coefficients is statistically different than 0, at better than a 0.05 level of significance. Accordingly, there is a strong correlation between the number of occurrences of the accounting words in blogs, the comments on the blogs and message boards, resulting in knowledge source clustering. This finding is important since it suggests that the analysis of the information contribution of either source independently (for example on the stock market) could reflect a correlated omitted variable.

Table 4
Number of occurrences of accounting terms in posts, blogs and message boards.

Accounting term	Posts	Blogs	Google	Lerman S&P 1500	Lerman non S&P
Earnings	858	305	12,928,103	65,218	34,679
Cash	2861	3274	78,685,721	33,160	31,012
Revenue	2453	826	29,018,691	22276	24,201
Dividend	100	82	1,287,983	20,927	8621
Buyback	19	2	133,289	18,287	9581
PE (price to earnings)	116	525	123,085	15,762	6368
Current report	1	49	21,684	13,294	12,645
EPS	254	46	168,706	12,158	5663
Asset	645	371	15,897,466	11,332	10,475
Cash flow	138	218	1,292,792	7566	5305
Periodic report	17	0	3,948	7539	10,407
Inventory	412	185	12,060,594	7283	3576
Expense	700	119	13,645,776	7130	6518
BS balance sheet	151	27	663,281	5747	3797
Impair	18	3	272,382	5393	3120
Accounting	386	823	17,425,414	4560	4080
Analyst	1263	470	18,405,449	4524	2101
Leverage	513	202	6,220,257	3974	3062
Profit	1459	2445	56,212,431	3904	3356
Book value	250	45	197,403	3536	3326
Stock option	83	127	169,257	3411	2060
Income	1840	3663	52,559,413	3151	3072
Guidance	820	593	14,919,618	3149	1842
Fair value	149	15	141,287	2716	1691
Liability	384	146	10,641,413	2662	1744
Lease	212	250	8,366,491	2215	1913
R&D	37	142	100,236	2042	2893
GAAP	147	15	212,729	1772	1606
Audit	161	260	6,446,494	1401	2854
CAPEX	9	2	99,389	1378	798
EBIT	0	1	84,393	1348	1120
Depreciate	11	3	161,602	1165	824
Financial instrument	13	0	56,561	1128	986
Unusual	1053	830	23,527,431	1034	732
Return on	149	212	1,487,300	1019	505
Pro forma	25	0	71,394	983	854
Restate	12	7	204,160	973	1023
Goodwill	157	51	1,352,587	884	558
Defer	101	5	460,944	805	1065
Income statement	92	11	86,255	752	686
Selling, general and administrative	0	0	3,123	717	481
Bad debt	259	177	204,133	611	268
Covenant	45	123	1,305,450	606	482
Equity	743	926	17,907,911	589	502
Receivable	3	33	267,147	534	686
Securitize	3	0	11,687	524	567
OBS (off balance sheet)	51	2	25,003	470	196
Accrue	35	0	327,506	459	567
Intangible	60	21	609,257	393	353
COGS (cost of goods sold)	14	1	24,631	345	217
PP&E (property, plant and equipment)	1	0	2,688	300	337
Discontinue	49	3	451,464	214	105
Market to book	0	62	52,551	211	121
Going concern	21	4	65,310	188	623
Cash flow Statement	10	8	30,225	178	145
Payable	20	12	914,599	168	162
AFS (available for sale)	41	71	6,413,402	94	156
Marketable securities	1	0	15,488	76	134
Minority interest	2	1	19,765	75	158
Pension	260	95	5,193,696	41	23
Contingent	176	14	954,560	34	22
HTM (held to maturity)	1	0	33,093	20	19
CI (comprehensive income)	47	10	6,251	16	34
MD&A (management discussion and analysis)	11	10	12,500	12	19

Two phrases "current" and "control" with substantial non accounting meaning are not included in the analysis.

7. Predicting sales from blog chatter

Blogs have also been used as the basis of generating information to predict sales. Gruhl et al. [9] apparently were among the first to examine the use of information from blogs as a basis to predict sales when they analyzed Amazon book sales. They found that peaks to discussions (chatter) in blogs were likely to be followed by sales

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

Table 5
Correlations between blogs, posts and message board for number of occurrences of accounting terms.

	Posts	Blogs	Google blogs	Lerman S&P 1500	Lerman non S&P
Posts					
Blogs	0.826				
Google blogs	0.919	0.942			
Lerman S&P 1500	0.455	0.262	0.352		
Lerman non S&P	0.619	0.382	0.501	0.934	

peaks. Mishne and Glance [25] extended that work by noting that the nature of the chatter (positive, negative, etc.) also helped in the prediction of sales. Mishne and Glance [24] also extended the results to investigate movie sales (box office revenues). Liu et al. [23] also investigated using blogs to predict movie box office revenues. Dhar and Chang [7], also reported in Computerworld [6], examined the relationship between blog activity and music sales, suggesting that when legitimate blog posts exceed a threshold of 40, before an album has been released, sales were about three times the average. More recently, Asur and Huberman [2] found that microblog chatter on Twitter also could be used to predict box office revenues.

There are a number of potential extensions to this research. First, so far the applications have been to classic Internet goods (books, movies and music), but there has not been much research into other settings. For example, does blog chatter help in the prediction of the sales of goods in other settings by, say, Procter & Gamble or Ford or Cisco. Second, as noted in Ref. [6], there is an implied causality "…it turns out that the volume of blog posts before a music album's release can significantly affect future album sales." However, it is not clear if blog activity is causing sales to increase or if blog activity is just a manifestation of interesting books, movies and music. Clearly, the two questions have different implications. If the blogging is causing the increase then firms need to take a proactive stance and make sure that some proactive blogs are written regarding their products. Third, so far, since blogs are relatively new and since this research has not been widely diffused, it is likely that researchers have examined blogs in a setting where the blog information was not being manipulated by a company to generate interest in the goods. However, now that there are headlines such as "Blog Chatter Can Triple Music Sales," it is likely that companies will take a proactive stance and generate blog chatter. This is likely to influence the ability of future researchers to examine these issues since having seen this relationship it is likely that firms will begin to proactively get chatter on movies, music and other products. Accordingly, another question is what would be the reaction of potential buyers if they knew that companies were trying to manipulate their buying choices with bogus chatter? The negative reaction might exceed the impact of all potential gains.

8. Corporate public image monitoring and forecasting

Recently, IBM [14,18] released what it referred to as a "public image monitoring solution." This system was designed to examine blogs and other externally available media so that the "public image" of an entity, such as a firm, could be monitored in real time. Such a system could be critical since, if public image could be monitored then firms could leverage that information and take appropriate actions to respond in real time to changes or concerns about their public image.

8.1. Motivations for firms to monitor and forecast public image

There are a number of different reasons that a company would want to monitor and forecast their public image. First, there is reason to believe that public image affects the stock price. For example, recently Goldman Sachs' public image was blamed for their stock price tumble (e.g., [16]). Further, one of the first firms to show interest in IBM's system to monitor public image was Morgan Stanley, a financial services firm [18]. If public image and stock price are related, then public image could be monitored and forecasted, and that could lead to being able to generate abnormal positive returns in the stock market. Second, public image can affect buying decisions. As noted by an IBM executive [18] "Organizations are struggling to understand what people are saying about them in public. That ends up having an impact on opinion and buying decisions." Accordingly, being able to monitor and forecast image may facilitate management of public image as it affects customer buying decisions. Third, an IBM release [14] claims that, "Companies are seeking new ways to better understand how they are viewed by customers, investors and other stakeholders who have an impact on their brand reputation… (public image monitoring and forecasting) can help clients track and analyze the pulse of the public in real-time, allowing organizations to be more responsive and deliver better service to their customers." Fourth, law suits also apparently impact public image. As a result, apparently companies including Wal-Mart have begun establishing a "war room" to keep track of what is being said about it. As noted by an IBM executive [5], "People are more likely to spout an opinion on a blog than call a company and complain. Organizations are starting to learn about what potential issues consumers are having with their companies and services. That market is difficult for companies to actively monitor."

8.2. Approach to monitoring and forecasting public image

How can "public image" be monitored and forecasted? One approach is to troll blogs for information that influences "public image," determine when a particular entity is the subject of a blog and if that blog is positive or negative. The monitoring aspect could then provide the user with those blogs that were either positive or negative so that the user could understand the nature of the findings. Then the numbers of positive and negative blogs would provide a gage as to public image. Time series of numbers of positive and numbers of negative public image blogs would be used to monitor and measure public image. Finally, extending that concept, using those same time series, companies could forecast into the future to begin to anticipate how the public image will evolve over time.

As noted by an IBM executive [5], "(The results are)…not going to be 100% 'noise' free. You make tradeoffs: do you want to get everything that's relevant or do you want to miss some things? The feedback I'm hearing from companies is, if you can get them 50% or 80% there that makes a huge impact. Instead, Public Image Monitoring…should be used as a supplement to actually reading blogs, he observed." However, as we noted above, firms may find that they are discussed in thousands or millions of blogs. If that is the case, then it would be impossible to read each item, and reading even more than a few would drive responsiveness down so that the results would not be available in real time. As a result, if firms are interested in what is being said about them in blogs then they will need to employ an automated system.

8.3. Testing the potential of "public image"

As a test of the concept of a system to investigate "public image," I analyzed the universe of blogs available in Google Blogs using some sample concepts that might influence corporate "public image." The results are summarized in Table 6.

First, in order to determine the extent to which the term "public image" was referenced in blogs I searched that term and found 263,164 pages. Unfortunately, the term "public image" apparently is also the name of a band. As a result, there is immediately a problem with identifying blogs that refer to "public image." In order to try and eliminate the impact of the band on the number of pages, I search to

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

Table 6
Number of google blog pages with terms (May 5, 2010).

Number of pages	Term 1	Term 2	Missing this term
263,164		Public image	
145,380		Public image	Music
80,205,228	Product		
445,356	Social responsibility		
192,770	Child labor		
171,138	Going green		
13,138	Product	Public image	Music
1,262	Social responsibility	Public image	Music
314	Going green	Public image	Music
199	Child labor	Public image	Music
6,083,989	Microsoft	Product	
1,061,429	IBM	Product	
541,221	Wal-Mart	Product	
29,086	Microsoft	Social responsibility	
12,144	IBM	Social responsibility	
9,289	Wal-Mart	Social responsibility	
38,819	Microsoft	Going green	
8,502	Wal-Mart	Going green	
6,195	IBM	Going green	
5,287	Microsoft	Child labor	
3,470	Wal-Mart	Child labor	
2,696	IBM	Child labor	
3,233	Microsoft	Public image	Music
542	Wal-Mart	Public image	Music
517	IBM	Public image	Music

exclude those pages with "Music" and got roughly one-half the number of pages. Although this approach is unlikely to fully mitigate the impact of a band with the same name as the search concept, it does illustrate some of the difficulties in trying to mine blogs. Second, in order to understand the order of magnitude of different components of the concept of "public image," I also searched on some other terms, including "social responsibility," "going green," "child labor," and "product" and found that "product" had substantially more than any of the other terms. Without specification of a company entity the results indicate that a very large number of blogs need to be examined. Third, of course this issue could be mitigated to a certain extent if an entity name such as "IBM" is included. As a test I included the number of blog pages for the individual components, when combined with a company entity, IBM, and phrase, "going green." As might be expected, such a joint search resulted in much smaller and more focused number of concepts being found. For example, "going green" showed up in 171,138 pages, but that same concept, with the term "public image" and missing the term "music" yielded 314 pages. Similarly, the concept "going green" showed up with Microsoft (38,819 pages), Wal-Mart (8,502 pages) and IBM (6,195 pages).

8.4. Limitations in the notion of "public image"

There are many components to "public image," of which some are listed in Table 6. However, a priori, it is not clear which components (e.g., "social responsibility" or "going green") should be included. Although there are some components to "public image" listed in Table 6, different companies are likely to have specific concerns and those would be captured in their conceptualization of "public image," consistent with their strategy. For example, consider a professional services firm that experienced law suits regarding a set of tax shelters that they devised. If public image were being monitored for them, the public's view of tax shelters might be important. On the other hand, since they are a professional services company, "child labor" would not be an issue for them.

Further, any set of components is not likely to be stationary. Instead, it is likely to change based on the environment and strategies that are executed by the firms.

In addition, some concepts by their very nature are positive or negative. For example, it would be difficult to imagine putting a positive spin on "child labor." Accordingly, for some concepts, simply counting numbers of blogs investigating the concept may be appropriate.

Further, it is not clear how "public image" as a summary measure or approach (such as the balanced scorecard) would be treated, particularly if some components were seen as containing more positive sentiment pages than other components. For example, suppose that "going green" was 80% positive, while "child labor" was 90% negative. What would a composite measure tell a company?

Finally, it is possible that if management ultimately is evaluated on the notions of "public image," they may take a direct role in generating data. For example, it is possible that managers could start blogs and make those blogs contain either positive or negative information, whatever is appropriate for the specific concept being measured. For this reason, it could be important for pre-specification of the specific blogs that were to be monitored.

9. Contributions and extensions

This paper has investigated some of the key issues in blog mining and some applications, generating a number of contributions to the existing literature on blog mining. First, rather than depending on bloggers to provide tagged mood information (which seems to be a disappearing information source), alternative tag information from readers (e.g., DElicious) can be used to assess the mood or sentiment of the blog. Second, different approaches can be used for content analysis of blogs to investigate the nature of the sentiment of the blog. This paper proposed looking for "declarations of mood" (e.g., "thumbs up") and information about the particular blogger and site to supplement other approaches, such as using generic positive and negative word dictionaries. Third, this paper suggested that domain specific terms be accounted for to improve the quality of the analysis. This paper found that generic positive and negative dictionaries had some limitations in describing negative behavior in the stock market. Fourth, this paper found that information from different knowledge sources, blogs and message boards appear highly correlated. In particular, the number of occurrences of accounting words in message boards is highly correlated with occurrences of those same accounting words in blogs and financial blogs. Fifth, as we noted above, there is a growing literature that links blog discussions to sales, at least for Internet goods. However, what is not clear is the causality link: Are the blogs just a manifestation of interesting books, movies and music or are the blogs actually influencing sales, or both? This paper questioned the causality and discussed some concerns associated with companies trying to directly generate blog chatter. Sixth, systems that use blog information to monitor and predict public image could be important to a firms for a number of different reasons. However, this paper found that public image terms, including "public image" can be difficult to generate searches because of the multiple concepts associated with the term. In addition, this paper argued that individual components, likely determined by strategy and environment should be directly monitored to capture the important issues associated with "public image".

However, there are a number of unsettled issues and extensions to the material in this paper. First, researchers have used news stories in the past as a source of text content [19]. However, additional research can be done to determine if blogs have a similar impact on key indices, such as the stock market and to what extent there is information content in blogs. Second, generating sentiment from text continues to provide research opportunities, although in some cases robust results are found simply using numbers of blogs. Unfortunately, the use of general opinion words as a means of establishing positive or negative context to a blog has some limitations. As an example, phrases such as "I love Microsoft bashing," found in some blogs, have both positive

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016

(love) and negative (bash) terms directly adjacent to the entity of concern "Microsoft." Accordingly, alternative approaches need to be examined and developed, breaking away from the individual word level of analysis. Additional analysis can drive toward the sentence or blog level following theorists such as Toulmin [32]. Further, generic positive and negative words were found to not be complete for a set of terms negatively describing a stock from input generated from a finance expert. Exactly how the domain might be more fully leveraged in determining sentiment remains an important research issue. Individual blogger consistency of opinion, blogger declaration and opinion dictionaries potentially could be coupled together to generate a stronger approach integrating and correlating information from multiple sources and using multiple approaches.

Third, future research could focus on the analysis of tags, such as those in DElicious. The extent to which tags are correlated with other tags and are redundant or the ability to cut concepts, such as "stupid" to "stupid" and "sad", and "stupid" and "funny" could be analyzed in more detail and used in generating an understanding of blog content.

Fourth, although sales of classic Internet goods have been shown to have blog chatter related sales, it is unclear if other goods, such as consumer goods, computer related goods (e.g., routers) or other categories of goods exhibit the same relationship. In addition, since the relationship between blog chatter and sales is so apparent, an important research issue is, how will firms respond? Further, if firms begin to generate their own blog chatter, what will be the reaction of customers when they find out some of the chatter is generated by the specific companies, and will the chatter actually lead to an increase in sales?

Fifth, definition and determination of what topics actually should be monitored as influencing "public image" are not clear. Further, how to integrate multiple component topics also is not clear. For example, should balanced scorecard-like approach be used to weigh the multiple factors? Also, to what extent are issues that are part of "public image" likely to be firm or industry or time dependent?

Finally, it is not clear if information contained in blogs relates back to financial measures, such as stock price. For example, the existence of a relationship between public image measures and stock price is not established, but could be investigated. Such an analysis could focus on concepts such as "going green" or other concepts not directly addressed here.

Acknowledgement

The author would like to thank the anonymous referees for their comments on an earlier version of this paper. The author would also like to acknowledge Jim Marsten's extensive comments on a more recent version of this paper.

References

[1] W. Antweiler, M.Z. Frank, Is all of that talk, just noise? The information content of Internet stock message boards, Journal of Finance 59 (3) (2004) 1259–1295.
[2] S. Asur, B. Huberman, Predicting the Future with Social Media, http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.5699v1.pdf 2010.
[3] G. Attardi, M. Simi, Blog Mining through Opinionated Words, http://trec.nist.gov/pubs/trec15/t15_proceedings.html 2006.
[4] K. Balog, M. Rijke, Decomposing Bloggers' Moods, WWW2006, Edinburgh, UK, May 22–26 2006.
[5] G. Clarke, IBM Targets Brand Conscious with Search, The Register, http://www.theregister.co.uk/2005/11/09/ibm_blog_brand_software/ Nov. 9 2005.
[6] Computerworld, Blog Chatter can Triple Music Sales, http://www.pcworld.idg.com.au/article/205589/blog_chatter_can_triple_music_sales/?fp=&fpid=&pf=1 Feb. 13 2008.
[7] V. Dhar, E. Chang, Does chatter matter? The Impact of User Generated Content on Music Sales, NYU Working Paper, 2007.
[8] J. Glionna, Workers Warned of Dangers at Toyota, Los Angeles Times, Monday, Mar. 8 2010, p. A1 and A10.
[9] D. Gruhl, R. Guha, R. Kumar, J. Novak, A. Tomlins, The Predictive Power of Online Chatter, in KDD 05 Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 78–87, ACM Press, New York, 2005.
[10] C. Haley, Blog Spotting with IBM, http://www.internetnews.com/xSP/article.php/3562116 Nov. 7 2005.
[11] T. Hampp, A. Lang, Semantic Search in WebSphere Integrator OmniFind Edition: The Case for Semantic Search, http://www.ibm.com/developerworks/data/library/techarticle/dm-0508lang/ August 5 2005.
[12] F. Hayek, The use of knowledge in society, The American Economic Review 35 (1945) 519–530.
[13] L. Hoang, J. Lee, Y. Song, H. Rim, Combining Local and Global Resources for Constructing an Error-Minimized Opinion Word Dictionary, PRICAI 2008, LNAI 5351, 688–697, Springer-Verlag, Berlin, 2008.
[14] IBM, IBM tracks Blogs, Web Content to Capture Buzz, Spot Trends Around Companies, Products and Marketing Campaigns, http://www-03.ibm.com/press/us/en/pressrelease/7961.wss Nov. 7 2005.
[15] S. Kim, E. Hovy, Determining the Sentiment of Opinions, Proceedings of the COLING, Geneva, 2004 http://www.isi.edu/natural-language/people/hovy/papers/04Coling-opinion-valences.pdf.
[16] V. Kumar, Goldman Shares Tumble as Image Takes Another Hit, http://www.dailyfinance.com/story/investing/goldman-shares-tumble-as-public-image-takes-another-hit/19460940/ Apr. 30 2010.
[17] L A Times, Magistrates Walk Out Over Changes, Los Angeles Times, Jan. 31 2010, p. A26.
[18] M. LaMonica, IBM to Analyze Digital Scuttlebutt, http://news.cnet.com/IBM-to-analyze-digital-scuttlebutt/2100-1012_3-5940339.html Nov. 8 2005.
[19] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, Language Models for Financial News Recommendations, Proceedings of the Ninth International Conference on Information and Knowledge Management, ACM Press, New York, NY, 2000, pp. 389–396.
[20] Leelefever, What are the Differences between Message Boards and Web Logs? http://www.commoncraft.com/what-are-differences-between-message-boards-and-weblogs-15 2004.
[21] A. Lerman, Individual Investor's Attention to Accounting Information: Message Board Discussions, unpublished paper, New York University (Jan. 2010).
[22] K. Liu, J. Zhao, NLPR at Multilingual Opinion Task in NTCIR7, Proceedings of the NTCIR-7 Workshop Meeting, Tokyo, Japan, Dec. 16–18 2008.
[23] Y. Liu, X. Huang, A. Ari, X. Yu, ARSA: A Sentiment Aware Model for Predicting Sales Performance Using Blogs, SIGIR '07, The Netherlands, July 23–27 2007.
[24] C. Mann, Spam + Blogs = Trouble, Wired 14 (9) (Sep. 2006).
[25] G. Mishne, N. Glance, Predicting Movie Sales from Blogger Sentiment, American Association for Artificial Intelligence, 2005.
[26] G. Mishne, M. Rijke, Moodviews: Tools for Blog Mood Analysis, American Association for Artificial Intelligence, 2006.
[27] I. Nonaka, The knowledge creating company, Harvard Business Review 69 (1991) 96–104.
[28] M. Polanyi, The Tacit Dimension, Routledge & Kegan Paul, London, UK, 1966.
[29] E. Ramstad, , Tuesday, Critic of Seoul Arrested, The Wall Street Journal, January 13 2009, p. 3.
[30] S. Sabherwal, S. Sarkar, Y. Zhang, Online Talk: Does it Matter, Managerial Finance 34 (6) (2008) 423–436.
[31] R. Tong, Detecting and Tracking Opinions in On-Line Discussions, UCB/SIMS Web Mining Workshop, 2001.
[32] S. Toulmin, The Uses of Argument, Cambridge University Press, Cambrdige England, 1964.
[33] R. Tumarkin, Internet Message Board Activity and Market Efficiency: A Case Study of the Internet Service Sector Using RagingBull.com, Financial Markets, Institutions and Instruments 11 (4) (2002) 313–335.
[34] V. Vara, Companies Mine Blogs for Market Research, Wall Street Journal, Dec. 3 2004, http://online.wsj.com/article/0,,SB110071998028976955-email,00.html.
[35] M. Xinfan, W. Houfeng, Detecting opinionated sentences by extracting context information, Proceedings of the NTCIR-7 Workshop Meeting, Tokyo, Japan, Dec. 16–18 2008.
[36] H. Yang, L. Si, J. Callan, Knowledge Transfer and Opinion Detection in the TREC 2006 Blog Track, http://trec.nist.gov/pubs/trec15/t15_proceedings.html 2006.

Daniel O'Leary is a Professor in the Marshall School of Business at the University of Southern California (USC), focusing on information systems, including artificial intelligence, enterprise resource planning systems and knowledge management systems. Dan is a full professor, with a joint appointment in accounting and IOM, and has been at USC since 1985. Dan received his Ph. D. from Case Western Reserve University. He received his MBA from the University of Michigan and his bachelor's degree from Bowling Green University (Ohio).

He is the former editor of IEEE *Intelligent Systems* and current editor of John Wiley's *Intelligent Systems in Accounting, Finance and Management*. His book, *Enterprise Resource Planning Systems*, published by Cambridge University Press, has been translated into both Chinese and Russian.

Please cite this article as: D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support Syst. (2011), doi:10.1016/j.dss.2011.01.016