Decision Support Systems 51 (2011) 821–830

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 1 February 2011

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