Quantifying Sentiment and Influence in Blogspaces

Peter Hui
Pacific Northwest National Laboratory
Richland, WA
peter.hui@pnl.gov

Michelle Gregory
Pacific Northwest National Laboratory
Richland, WA
michelle@pnl.gov

ABSTRACT

The weblog, or blog, has become a popular form of social media, through which authors can write posts, which can in turn generate feedback in the form of user comments. When considered in totality, a collection of blogs can thus be viewed as a sort of informal collection of mass sentiment and opinion. An obvious topic of interest might be to mine this collection to obtain some gauge of public sentiment over the wide variety of topics contained therein. However, the sheer size of the so-called *blogosphere*, combined with the fact that the subjects of posts can vary over a practically limitless number of topics poses some serious challenges when any meaningful analysis is attempted. Namely, the fact that largely anyone with access to the Internet can author their own blog, raises the serious issue of credibility— should some blogs be considered to be more *influential* than others, and consequently, when gauging sentiment with respect to a topic, should some blogs be weighted more heavily than others? In addition, as new posts and comments can be made on almost a constant basis, any blog analysis algorithm must be able to handle such updates efficiently. In this paper, we give a formalization of the blog model. We give formal methods of quantifying sentiment and influence with respect to a hierarchy of topics, with the specific aim of facilitating the computation of a per-topic, influence-weighted sentiment measure. Finally, as efficiency is a specific end-goal, we give upper bounds on the time required to update these values with new posts, showing that our analysis and algorithms are scalable.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles

General Terms

Algorithms

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

Keywords

Social Media Analytics, Influence, Sentiment, Formal Methods

1. INTRODUCTION

In recent years, the widespread presence of web-based outlets gives the opportunity for anyone with access to the Internet to contribute their thoughts and opinions to a potentially world-wide audience. The existence of myriad social networking services and other forms of social media provide ample outlets for such communication. Specifically, the weblog (henceforth, blog) has become a common and accessible venue for people to post information on a wide variety of topics, and such, the so-called *blogosphere*— the collective aggregation of blogs across the web— can be thought of as a large cross-section of mass opinion on a myriad of topics. Moreover, mining this large body of collective thoughts can prove to be one potentially powerful method of gauging public sentiment on any number of topical areas.

The fact that largely anyone can post their opinions on a blog has both advantages and disadvantages; on the positive side, it serves as a widely accessible outlet for publishing one's thoughts and ideas. On the negative side, however, the fact that any author may host a blog raises the issue of credibility— one can easily argue that in many cases, not all authors should be considered to be equally credible; with respect to a given topic, some authors should be considered to be more credible than others. For example, consider two blogs B and C, where both post a large percentage of their posts on the topic of the world economy. Suppose that B is authored by a widely acknowledged expert in the field, having a large audience, and which generates a large amount of feedback in the form of post comments. On the other hand, C is authored by an undergraduate student who is just starting to learn about this topic, and uses this blog as a venue to express his opinions on this field. He has a small audience consisting of family, friends, and fellow students, and his posts generate few comments. Intuitively, one would consider B to be much more *influential* than C, and consequently any measure of sentiment between these two blogs should be weighted more heavily in favor of B, and this weighting is ultimately what we intend for our framework to capture.

Integrating these concepts, we propose a novel method for analyzing the contents of a collection of blogs with the specific goal of gauging the sentiment of a wide variety of topics. Specifically, we aim to capture the aforementioned notion of weighting the credibility of a blog, giving higher weight to

blogs deemed to be more *influential*, thus giving the ability to compute an *influence-weighted* sentiment quantification. In order to do so, we propose methods for explicitly quantifying both of these arguably subjective properties with several end goals in mind.

In no particular order, our first goal is for our analysis to be able to quantify influence and sentiment with respect to a wide variety of topics; as the number of topics contained within a single blog, let alone a large collection of blogs, can be almost unlimited, a measure of "influence" of a particular blog is arguably of little use unless it is with respect to a certain topic. For example, if we are interested in gauging public sentiment in the area of the global economy, then a blog that is otherwise deemed to be highly influential is of little use if the majority of posts contained within it deal with an irrelavent subject area— basket weaving, for instance. Specifically, we envision the collection of topics to form a hierarchy. This gives many benefits. For one, such a hierarchical arrangement poses a very natural characterization of a use-case for a practical implementation. Namely, when performing such an analysis, one is likely to be interested in gauging influence and sentiment with respect to a closely related and organized set of topics, rather then an ad-hoc collection of unrelated ones. Secondly, this allows for the user to refine analysis with increasing granularity. For example, consider the broad topic of "The economy of Country X", under which lies the subtopics of taxation, job growth, and the stock market. Supposing that an analysis of the parent topic shows a generally negative sentiment, this hierarchical arrangement of topics allows one to further analyze which of the subtopics specifically is contributing to this overall negative sentiment. An ad-hoc, unorganized topical arrangement would not allow such a refinement of analysis. An example of a topic hierarchy is shown in Figure 1.

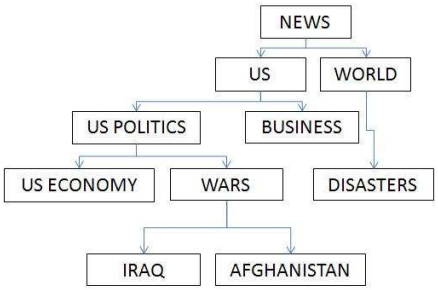

Figure 1: Example of a topic hierarchy over news topics. Any post on the topic of the "US Economy" would also be considered to be a post on the topic of "US Politics", "US (news)", and "News" as well.

The second goal is to quantify these values in a manner which reflects an intuitive notion of what it means for a blog to be considered to be *influential*. At a high level, we propose that with respect to a given topic, an *influential* blog is one which:

- has a non-trivial number of followers,
- generates a non-trivial amount of user feedback, in the form of comments on posts, and
- has a large proportion of posts on the topic being analyzed

Admittedly, the phrase *non-trivial* is a subjective one; we formally quantify this notion (using terms $\text{In}(B)$ and $\tau(B, \mathcal{T})$, respectively) in Section 3.3.

In addition, if one blog is deemed to be influential with respect to the topic being analyzed, and it follows (in the sense that one blog's author follows, or "subscribes" to other blogs' posts) a second blog which publishes a large proportion of posts in this same topic, one would intuitively expect this to speak positively about the latter blog, indicating that the latter blog should be considered to be influential in this topic area as well. We elaborate more on these criteria in Section 3.3.1.

The third goal in our analysis is to explicitly ensure the ability to efficiently analyze the entire blogspace with respect to each topic in the topic set. This in turn poses its own unique set of challenges. For instance, one must be able to quantify sentiment between any distinct topics in the topic set without expensive recomputation when "switching views" between topics in this sense. In addition, since we assume that topics form a hierarchy, the analysis must satisfy the expected compositional properties— namely that the measure of sentiment for one topic must be appropriately reflective of those of its children.

Fourthly, as blogs by their very nature are dynamic and rapidly evolving with new posts and comments processed on a very frequent basis, we require the ability to process incremental updates efficiently. Specifically, the addition of a new comment or post should be able to be processed incrementally, rather than having to recompute all values over the entire blogspace and topic hierarchy.

A fifth overarching goal for our framework is the ability for our analysis to capture the intuition, as discussed above, that when quantifying sentiment in a blogspace, that blogs written by more *influential* authors should be weighted more heavily than those deemed less so. To achieve this last goal, we separately define the quantities of influence and sentiment, parameterized on topics, which in turn allows the computation of an influence-weighted sentiment over the blogspace on a per-topic basis.

The model for our work is that of the the *blog* model. Namely, we consider a collection of *blogs*, each of which consists of a series of posts, each of which in turn has a (possibly empty) set of comments. In addition, blogs can subscribe to, or *follow* other blogs; although in reality, the blogs' *authors* are actually following other blogs, for the purpose of our analysis, we identify blogs with their authors. In addition, we assume the existence of functions which extract the topic of a post, as well as those which analyze the sentiment of a comment as a value in the range $[0, 1]$. Treating these components as abstract components is not intended to trivialize these important and difficult tasks, but rather to allow us to focus on the other facets of this overall challenging problem. We discuss this approach more in Section 3.2. Figure 2 depicts this model visually.

The context for our work is ultimately a software framework for carrying out the analysis described in this paper. Specifically, we ultimately envision a framework into which a collection of blogs can be loaded, and an influence-weighted sentiment analysis can be computed with respect to a given set of topics. While this is the end-goal, we have not yet developed such a framework, but rather, this paper is instead devoted to establishing the mathematical formalisms which will underpin this framework. To this end, we present

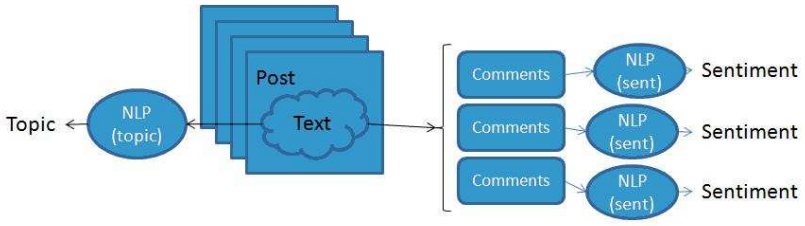

Figure 2: Our model of a blog. Each blog consists of a series of posts. Each post consists of a body, as well as a series of comments. Our analysis assumes the existence of natural language processing (NLP) components which extract the topic of the blog's post (given the topic hierarchy), and the overall sentiment of each comment as a numerical value in the range [0,1]. Our analysis treats these NLP components as black boxes, abstracting away the details of these complex components, allowing us to focus on the remaining analysis.

the theoretical foundations for our analysis, leaving actual implementation and the associated details for future work.

The main contributions of this paper are as follows: We give formal methods of quantifying influence and sentiment in a blogspace, specifically with respect to a hierarchy of topics. As stated above, the ability to process updates efficiently is of specific importance, and as such, we give theorems giving explicit time bounds for computing and updating these sentiment values.

The rest of the paper proceeds as follows: in Section 2, we discuss related work in the field. In Section 3, we present our formal model of a blogspace, we define methods to quantify influence and sentiment in this model, and we give results on time bounds for computing and updating these values. In Section 4, we discuss practical implications, consequences, and technicalities related to our model. We conclude with Section 5.

2. RELATED WORK

There has been much previous work aimed at studying social media, ranging from broad, general-purpose social network analysis tools [24], to structural analysis of social networks [27].

The term "social media" itself is a broad term, lending itself to a wide variety of models, many of which have been the areas of recent research. While our work focuses specifically on the blog model, there are others as well. For instance, the *microblogging* ("Twitter") model has been studied to determine the very nature of its usage [11]. Broad social online communities have been studied in the context of predicting human behaviour as well— see, for instance [26].

The area of collections of blogs specifically has been the area of much recent research as well, including work on methods of distilling the topical content of a blog [17], and *topical* clustering to analyze topic trending within a blog [23]. Another related topic is the area of inferring or extracting broad clusters of topics or communities from a collection of blogs [2, 4, 13, 15], as well as temporal analysis of social media in general [4, 19, 30]. In terms of other, non-structural, non-topical types of analysis, there has been research in detection of so-called "spam" blogs (splogs) [16, 29]. He, et. al. have done recent work in novel, statistical approaches to extracting opinionated posts from a blog [10].

In [3], Chi, Tseng, and Tatemura detect temporal trending within aggregates of blogs on related topics, reflecting changes in the blog over time.

In [25], Song, Chi, Hino, and Tseng give an algorithm and some experimental results aimed at identifying "opinion leaders" in a blogspace using a measure of a blog's influence or opinion leadership based on the *novelty* or originality of the information contained within the blog, giving higher weight to those blogs containing original material (versus those having a higher percentage of "reposted" material).

Finally, there has been work as well studying ways of integrating the *Pagerank* [1] algorithm with a set of topics. See, for instance, [9, 21].

Our work is complementary to, and distinguished from, each of the previous works. Namely, our work is an attempt at quantifying sentiment, weighted by influence. In addition, our analysis is with respect to a given hierarchy of topics, with a specific emphasis on efficiency of updates as a goal.

3. MODEL

We begin by defining our model, which assumes a blogspace as a collection of blogs, each of which consists of a series of posts. Each post consists of a body of text and a series of comments made on that post. Figure 2 gives a graphical depiction of this model. We assume the existence of natural language processing components, depicted in Figure 2 as *NLP (topic)* and *NLP (sent)*, which process the post's body and comments, extracting the topic of the post, and a measure of the sentiment of each comment, respectively. We discuss this decision more in Section 3.2. As an example of the former case, this NLP component may analyze the text of a post and conclude that the topic of the post is a recent rally in the stock market. As an example of the latter case, the NLP component analyzes each comment, returning a measure of the sentiment as a real number in the range e.g. $[0, 1]$; thus a particularly negative comment might receive a sentiment score of 1.0×10^{-2}, while a particularly positive comment might receive a score of 0.95. In addition, we assume comments have been sanitized, in that meaningless comments such as advertisements (in colloquial terms, *SPAM*) have been removed, leaving only meaningful comments for use in our analysis. This can easily be achieved in our actual implementation, for instance, through the use of existing filtering software.

3.1 Preliminaries

Before beginning our formal analysis, we give a few mathematical preliminaries. Blogs are ranged over by metavari-

able B. A set of blogs is called a *blogspace*, and is denoted as a set $\bar{B}$. Each blog consists of a vector of posts, and for a blog B, the function $\textsc{Pst}(B)$ returns this post vector. Individual posts are denoted using metavariables p and q, and a vector of posts is denoted as a vector $\vec{p}$. Topics in the hierarchy are denoted using metavariable $\mathcal{T}$, and the function $\textsc{Topic}(p)$ gives the topic in the hierarchy of the post p. A topic hierarchy is denoted as a pair $(\bar{\mathcal{T}}, <)$, where $\bar{\mathcal{T}}$ is the set of topics, and $<: \bar{\mathcal{T}} \rightarrow \bar{\mathcal{T}}$ is the subtopic relation, mapping topics to topics, which defines the parent-child relationships amongst topics in the hierarchy. Without loss of generality, we assume that the relation $<$ is well-formed; that is, that the topic hierarchy forms a tree. We write $\preceq$ as the reflexive, transitive closure of $<$ and $\lesssim$ as the transitive closure of $<$.

In addition, we extend the function $\textsc{Pst}(B)$ to allow specification of only the posts relevant to a given topic $\mathcal{T}$:

$$\textsc{Pst}(B, \mathcal{T}) = \{p \in \textsc{Pst}(B) \mid \textsc{Topic}(p) \preceq \mathcal{T}\} \quad (1)$$

$$\textsc{Pst}_{\lesssim}(B, \mathcal{T}) = \{p \in \textsc{Pst}(B) \mid \textsc{Topic}(p) \lesssim \mathcal{T}\} \quad (2)$$

$$\textsc{Pst}_{=}(B, \mathcal{T}) = \textsc{Pst}(B, \mathcal{T}) \setminus \textsc{Pst}_{\lesssim}(B, \mathcal{T}) \quad (3)$$

For a set S, we write $|S|$ to denote the cardinality of S. We also define a corresponding function $\#$ which gives the cardinality of the respective sets:

$$\#(B, \mathcal{T}) = |\textsc{Pst}(B, \mathcal{T})|$$
$$\#_{\lesssim}(B, \mathcal{T}) = |\textsc{Pst}_{\lesssim}(B, \mathcal{T})|$$
$$\#_{=}(B, \mathcal{T}) = |\textsc{Pst}_{=}(B, \mathcal{T})|$$

Each post consists of a vector of *comments*, and for a post p, the function $\textsc{Comm}(p)$ returns this comment vector.

3.2 Sentiment

We begin this section by explicitly noting that our sentiment analysis is parameterized in part on two functions. First, we assume the existence of a function $\textsc{Topic}(p)$ which analyzes a post and extracts the topic of the post, given the list of topics in the hierarchy as potential candidates. We fully realize that this itself is not a trivial task, and is in fact the area of active research as well [6, 7, 12, 18, 31]. Our intent in doing so is not to trivialize this task. Rather, our intent is to abstract away the details of this complex task by treating it as a sort of "black box", allowing us to concentrate our analysis on the facets of the problem distinct from that task, similar to the manner in which oracle Turing machines [8] are parameterized on their respective oracles. In addition, as this area is itself the topic of much research, explicitly decoupling our analysis from this functional component will allow us to treat this piece as interchangeable, allowing us to try different approaches to this problem in our future implementation.

Similarly, we assume the existence of a function $\sigma(c)$ which calculates the sentiment (e.g. [5, 14, 20, 22, 28]) of a text string as a value in the range $[0, 1]$. We compute the *sentiment* of a post p as the mean of the sentiment values of the comments in p's comment vector:

$$\sigma(p) = \sum_{c \in \textsc{Comm}(p)} \sigma(c) \cdot \frac{1}{|\textsc{Comm}(p)|} \quad (4)$$

We extend this function to compute the sentiment of a blog B with respect to a topic $\mathcal{T}$ as one would expect—namely, the mean of the sentiment values of the posts in B's post vector whose topic is $\mathcal{T}$:

$$\sigma(B, \mathcal{T}) = \sum_{p \in \textsc{Pst}(B, \mathcal{T})} \sigma(p) \cdot \frac{1}{\#(B, \mathcal{T})} \quad (5)$$

$$= \frac{\sum_{p \in \textsc{Pst}(B, \mathcal{T})} \sigma(p)}{\#(B, \mathcal{T})} \quad (6)$$

$$= \frac{\sum_{p \in \textsc{Pst}_{=}(B, \mathcal{T})} \sigma(p) + \sum_{p \in \textsc{Pst}_{\lesssim}(B, \mathcal{T})} \sigma(p)}{\#_{=}(B, \mathcal{T}) + \#_{\lesssim}(B, \mathcal{T})} \quad (7)$$

where the last step follows from substitution with equations 1–3.

We extend the function further to compute the sentiment of a blogspace $\bar{B}$ with respect to some topic $\mathcal{T}$ as the average of the sentiment of each of $\bar{B}$'s component blogs with respect to $\mathcal{T}$; this time, however, the average is weighted using a simple weighting function $w(B, \mathcal{T})$ (discussed below and given in Equation 11):

$$\sigma(\bar{B}, \mathcal{T}) = \sum_{B \in \bar{B}} \sigma(B, \mathcal{T}) \cdot w(B, \mathcal{T}) \quad (8)$$

The intent is to capture the notion that, within a blog, there can be posts on many different topics; for instance, one author might write 80% of her posts on the topic of data mining, while the other 20% of her posts might deal with the subject of the global economy. Then when computing the sentiment over the entire blogspace with respect to the topic of the global economy, one would expect this particular blog to contribute proportionally less to the sentiment value than one in which 95% of the posts deal with this same topic (namely, .20 vs. .95, respectively).

Thus, the weighting function is defined in terms of a characteristic function $\chi(p, \mathcal{T})$, which evaluates to 1 if the topic of post p is a subtopic of $\mathcal{T}$, and 0 otherwise:

$$\chi(p, \mathcal{T}) = \begin{cases} 1 & \text{if } \textsc{Topic}(p) \preceq \mathcal{T} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

We define an intermediate weighting function $w'(B, \mathcal{T})$, equal to the proportion of posts in blog B whose topic is a subtopic of $\mathcal{T}$:

$$w'(B, \mathcal{T}) = \frac{\sum_{p \in \textsc{Pst}(B)} \chi(p, \mathcal{T})}{|\textsc{Pst}(B)|} \quad (10)$$

We then normalize the intermediate weights so that over the entire blogspace, the sum of all intermediate weights with respect to topic $\mathcal{T}$ sum to 1, which in turn enables us to treat Equation 8 as a true weighted sum:

$$w(B, \mathcal{T}) = \frac{w'(B, \mathcal{T})}{\sum_{C \in \bar{B}} w'(C, \mathcal{T})} \quad (11)$$

We are ultimately interested not only in the value $\sigma(\bar{B}, \mathcal{T})$ for a single topic $\mathcal{T}$ across the blogspace $\bar{B}$, but rather the *sentiment vector* of values $\langle \sigma(\bar{B}, \mathcal{T}_1), ..., \sigma(\bar{B}, \mathcal{T}_n) \rangle$ for all $\mathcal{T}_i$ in the topic hierarchy. For a topic hierarchy $\bar{\mathcal{T}}$, we denote this vector as $\varsigma(\bar{B}, \bar{\mathcal{T}})$. Note that while the topic hierarchy is strictly defined as a pair $(\bar{\mathcal{T}}, <)$, we elide the relation $<$ out of succinctness. Similarly, when referring to a specific vector $\varsigma(B, \bar{\mathcal{T}})$, we will sometimes elide the B and $\bar{\mathcal{T}}$ as well:

Definition 1. For a topic hierarchy $\bar{\mathcal{T}}$ and blog B,

$$\varsigma(B, \bar{\mathcal{T}}) = \langle \sigma(B, \mathcal{T}_1), ..., \sigma(B, \mathcal{T}_n) \rangle$$

for all $\mathcal{T}_1, ..., \mathcal{T}_n \in \bar{\mathcal{T}}$. In addition, for a given vector ς, $\varsigma[\mathcal{T}]$ denotes the element $\sigma(B, \mathcal{T}')$ of ς where $\mathcal{T}' = \mathcal{T}$. Similarly, for a blogspace $\bar{B}$,

$$\varsigma(\bar{B}, \bar{\mathcal{T}}) = \langle \sigma(\bar{B}, \mathcal{T}_1), ..., \sigma(\bar{B}, \mathcal{T}_n) \rangle$$

This brings us to the main results of this section. As noted earlier, the ability to efficiently update the vector $\varsigma(\bar{B}, \bar{\mathcal{T}})$ with new posts is of importance, due to the rapidly evolving nature of blogs. We prove that our formulation lends itself naturally to such efficient updates by giving an algorithm which performs such an update, and show that it runs with reasonable efficiency. The heart of the following theorem lies in Algorithm 1, which we briefly describe, and for which a detailed analysis can be found in the proof. The algorithm takes a sentiment vector $\varsigma(B, \bar{\mathcal{T}})$ for a single blog and topic hierarchy, and updates the vector with respect to a new post q. Briefly, we maintain a tree structure which mirrors that of the topic hierarchy. The nodes of this tree correspond to the nodes of the hierarchy. With each node corresponding with a topic $\mathcal{T}$, we associate the number of posts whose topic is exactly $\mathcal{T}$ (denoted using the function $\#_{=}(\mathcal{T}, B)$), along with the value $\varsigma[\mathcal{T}]$ of that topic. To update the vector with a new post, we simply compute the topic of the new post, update the values in the node corresponding to that topic's node, and propagate these values upwards. Assuming a balanced topic hierarchy in the average case gives us an upper bound on the number of propagations upwards, and putting these numbers together gives us the desired bounds.

The correctness of the algorithm, for which details can be found in the proof, are due to the formulation of the sentiment values (Equations 1–11), which are carefully formulated in such a manner as to guarantee the requisite compositional properties.

LEMMA 1. *Assume that $\sigma(c)$ is computable in time $O(f(|c|))$ for some function f, where $|c|$ denotes the length of comment c, and assume that the topic of a post p can be computed in time $g(|p|)$, where $|p|$ denotes the length of post p. Then in the average case, a sentiment vector $\varsigma(B, \bar{\mathcal{T}})$ for a blog B can be updated with a new post p in time*

$$O\big(g(q) + [m \cdot f(n) \cdot \log(t)]\big)$$

where m is the number of comments on the new post p, n is the length of the longest such comment, q denotes the length of the new post, and t is the number of topics in the hierarchy.

PROOF. *We prove first that Algorithm 1 correctly updates a sentiment vector $\varsigma(B, \bar{\mathcal{T}})$ with a new post p, and secondly that it runs in the stated time.*

Correctness: *We maintain the following loop invariant: prior to each iteration of the loop (line 1), $\varsigma[\mathcal{T}']$ holds the value $\sigma(B, \mathcal{T})$ for all $\mathcal{T}' \precneqq \mathcal{T}$.*

- Initialization: *We start by showing that the invariant holds before the first iteration of the loop. By definition, $\varsigma[\mathcal{T}]$ holds the value $\sigma(B, \mathcal{T})$ for each $\mathcal{T}$ in the hierarchy, so the invariant is trivially true before the first iteration.*
- Maintenance: *Next, we show that the invariant holds after each iteration of the loop. At line 1, total holds the value $\#_{=}(B, \mathcal{T}) + \#_{\precneqq}(B, \mathcal{T})$, which, by Equation 6, means that $\sigma' \cdot total = \sum_{p \in \text{PST}(B, \mathcal{T})} \sigma(p)$. Therefore, the value computed at line 1 is*

$$\frac{\sum_{p \in \text{PST}(B, \mathcal{T})} \sigma(p) + \sigma(q)}{\#_{=}(B, \mathcal{T}) + \#_{\precneqq}(B, \mathcal{T}) + 1}$$

  which, by Equation 7, equals $\sigma(B, \mathcal{T})$ with new post q added.
- Termination*: The loop invariant states that $\varsigma(B, \bar{\mathcal{T}})$ holds the value $\sigma(B, \mathcal{T}')$ for all $\mathcal{T}' \precneqq \mathcal{T}$. At termination, $\mathcal{T}$ is the topmost topic in the hierarchy, which means that $\varsigma(B, \bar{\mathcal{T}})$ holds the correct values for all topics in the hierarchy, proving that the algorithm is correct.*

Time bounds: *Line 1 runs in time $g(q)$, and the computation of $\sigma(q)$ in line 1 takes time $O(m \cdot f(n))$ to compute the sentiment for each of q's comments. In the average case, we would expect the topic hierarchy to form a balanced tree with depth $O(\log(t))$ (and hence the outer loop (lines 1–1) runs $O(\log(t))$ times), and a constant branching factor for each node in the hierarchy. This allows us to consider lines 5–7 as a constant factor, as are the remaining lines in the loop, giving the desired time bounds.* □

Algorithm 1 Updates $\varsigma(B, \bar{\mathcal{T}})$ with a new post q

```
1: T = TOPIC(q)
2:
3: repeat
4:    total ⇐ #=(B, T)
5:    for T' where T' < T do
6:       total ⇐ total + #=(B, T')
7:    end for
8:    σ' ⇐ ς[T]
9:    #=(B, T) ⇐ #=(B, T)+1 {account for the new post}
10:   ς[T] ⇐ (σ'·total+σ(q)) / #=(B, T')
11:   T ⇐ parent(T)
12: until T == null
```

While the previous lemma gave an upper bound for updating the sentiment vector for a single blog B, this brings us to the main theorem, which gives an upper bound for updating the sentiment vector for an entire blogspace $\bar{B}$. These bounds are derived directly from those given in Lemma 1.

THEOREM 2. *Assume a blogspace $\bar{B}$, and f, g, m, n, q and t as in Lemma 1. Then in the average case, a sentiment vector $\varsigma(\bar{B}, \bar{\mathcal{T}})$ for blogspace $\bar{B}$ can be updated with a new post in time*

$$O\big(g(q) + \log(t) \cdot (m \cdot f(n) + |\bar{B}|)\big)$$

PROOF. *Let B be the blog in which the new post q was made. To derive an algorithm for updating the sentiment vector for a blogspace $\bar{B}$, we maintain a sentiment vector for each blog in the blogspace. Alongside each such vector, we maintain a vector of blog weights $W' = \langle w'(B, \mathcal{T}_1), ..., w'(B, \mathcal{T}_n) \rangle$, where w' is defined as in Equation 10. To process the update with new post q, we first update the sentiment vector*

for blog B, which, by Lemma 1, can be performed in time $O(g(q) + [m \cdot f(n) \cdot \log(t)])$. Next, we update the weight vector W'. Since, in the average case, we again assume a balanced topic hierarchy with height $\log t$, this step takes $O(\log t)$ steps. Next, we need to compute the normalized weight vector $W = \langle w(B, \mathcal{T}_1), ..., w(B, \mathcal{T}_n)\rangle$, where w is defined as in Equation 11. Doing so requires the computation of Equation 11, which in turn requires summation over each blog for each topic updated in the previous step. Since the previous step updated $\Theta(\log t)$ topics, this step takes $O(\log t) \cdot |\bar{B}|$ steps. Finally, we need to compute the weighted sum given by Equation 8, which again requires summation over all blogs for each topic updated in the previous step, requiring another $O(\log t) \cdot |\bar{B}|$ steps. At this point, we are done. The total number of steps required is thus

$O(g(q)) + [m \cdot f(n) \cdot \log(t)]) + \Theta(\log t) + 2 \cdot \Theta(\log(t) \cdot |\bar{B}|) =$
$O(g(q)) + [m \cdot f(n) \cdot \log(t)]) + 2 \cdot \Theta(\log(t) \cdot |\bar{B}|) =$
$O(g(q) + \log(t) \cdot (m \cdot f(n) + |\bar{B}|))$ □

3.3 Influence

In this section, we describe a method of computing *influence* within a blogspace, with respect to some topic $\mathcal{T}$. Our algorithm is influenced heavily by existing web-ranking algorithms, specifically that of Pagerank [1].

3.3.1 Criteria

One inherent difficulty in developing any algorithm for quantifying influence is that influence itself is, in a sense, a very subjective measure, and as such, any algorithm which computes a measure of this quality will need to first define the criteria used in computing this value. In this vein, the intent behind our algorithm for computing influence is to capture a very *natural* characterization of what it means for a blogger to be considered to be influential within the blogspace, with respect to a given topic. To this end, we aim to capture the following notions:

1. **Number of followers**— To be considered to be influential, a blogger must have a sizeable number of followers; no matter what a blogger publishes, if he has few or no followers, then he cannot be considered to be influential in any topic, as this implies that few people are listening to what he has to say.

2. **Relevancy**— To be considered to be influential with respect to a topic $\mathcal{T}$, a blogger must publish some fraction of her posts related to that topic; a blogger may be considered to be a widely revered, influential expert in the area of database systems, but if she never publishes any posts in the area of computer security, her blog cannot be considered to be influential in that topic.

3. **Comments**— This criterion captures the notion that an influential blogger will have more comments made on his posts than one who is less so; if none of the blog's readers are inclined to leave comments on the posts, it is hard to argue that the blog is an influential one.

4. **Followers**— This criterion captures the notion that if a blogger, Alice, who is deemed to be influential (with respect to some topic $\mathcal{T}$) follows a second author, Bob's, blog, and a large percentage of Bob's posts are on the topic $\mathcal{T}$ as well, then this should be considered to be a good indicator that Bob should be considered to be an influential author as well. Note that in this case, Alice may not necessarily view Bob's blog *favorably*. This is just an indication that her blog is of significant interest, be it in a positive or negative light; influence, not opinion, is the criteria being measured here.

   For example, suppose that Alice is considered to be an expert in the field of data mining, and in her blog, 95% of her posts are on this topic (*relevancy*). She has many followers (*number of followers*), and she has several comments on her posts, indicating, for instance, that her posts are insightful, meaningful, and generate feedback from her followers (*comments*). Now, suppose Bob also writes a high percentage of his posts on data mining. If Alice follows Bob's blog, then one might naturally think that this should be considered a sign that Alice trusts and respects Bob's views on this topic to be relevant enough to merit following his blog, and consequently that Bob should be considered to be influential in this area as well. However, if Bob doesn't post at all on this topic, or if he writes only a small fraction of posts on this topic, then if Alice is following Bob's blog, it is almost certainly not due to his expertise on the topic of data mining, and one cannot make any conclusions on his influence based on Alice's following of his blog.

We now give some definitions used to quantify these criteria:

For a blog B, we write $\text{In}(B)$ to denote the set of blogs which follow B, and we write $\text{Out}(B)$ to denote the set set of blogs which B follows. These functions will aid in quantifying criteria 1 and 4.

For a blog B and topic $\mathcal{T}$, $\pi(B, \mathcal{T})$ denotes the proportion of posts in B whose topic is $\mathcal{T}$. That is,

$$\pi(B, \mathcal{T}) = \frac{\sum_{p \in \text{Pst}(B)} \chi(p, \mathcal{T})}{|\text{Pst}(B)|} \tag{12}$$

Note that this is the same quantity as defined in Equation 10. However, since these two functions are used for different purposes in their respective contexts, we refer to this quantity as $\pi(B, \mathcal{T})$ in this context to avoid any confusion. This quantity will aid in quantifying criteria 2.

Finally, to quantify criteria 3, we need the ability to capture the notion of a *sufficient number of comments*. To this end, we take the approach of assuming some base threshold of the number of comments on a post. Above this threshold, we assume that the post has generated sufficient feedback or input from the post's readers to be considered to have a sufficient number of comments. We then capture this notion using a function $\tau(B, \mathcal{T})$ for topic $\mathcal{T}$ and blog B. The idea behind this function is that each post in $\text{Pst}(B)$ contributes some fraction of $\frac{1}{|\text{Pst}(B,\mathcal{T})|}$, so that in the best case, if all posts in the blog exceed the comment threshold, each post contributes exactly $\frac{1}{|\text{Pst}(B,\mathcal{T})|}$, and the function sums to 1. In the worst case, where all posts have zero comments, each post contributes a weight of 0, and the function in turn sums to 0. All other cases lie somewhere in between 0 and 1, depending on the number of comments left on the component posts. The function $\tau(B, \mathcal{T})$ is defined as follows, parameterized on τ_{max}, the base comment threshold:

$$\tau(B,\mathcal{T}) = \frac{\sum_{p \in \text{PST}(B,\mathcal{T})} \min\left(\frac{|\text{COMM}(p)|}{\tau_{max}}, 1\right)}{|\text{PST}(B,\mathcal{T})|} \quad (13)$$

EXAMPLE 3. *Consider a blog B containing 20 posts whose topic is $\mathcal{T}$. Let $\tau_{max} = 10$ comments; posts containing greater than 10 comments are deemed to have generated a sufficient number of comments to be considered to have been an influential post. Suppose 12 of these posts have greater than 10 comments, 3 of these posts have 6 comments, while the remaining 5 posts have 0 comments. Then*

$$\tau(B,\mathcal{T}) = \frac{(12 \cdot 1.0) + (3 \cdot 0.6) + (5 \cdot 0)}{20} = 0.609$$

EXAMPLE 4. *Consider the blog from Example 3. Suppose all 20 comments have greater than τ_{max} comments. Then*

$$\tau(B,\mathcal{T}) = \frac{20 \cdot 1}{20} = 1.0$$

3.3.2 Algorithm

Our proposed influence measurement algorithm is a variant of that of [1] (*Pagerank*), and can be viewed as a customization of that algorithm to suit our needs. We chose this particular algorithm for a few reasons. First, our influence criteria #4— the notion that, with respect to a given topic, if an influential blog follows a blog who itself posts frequently on the same topic, this second blog should as well be considered influential— is aligned very closely with the intuition captured by *Pagerank*. Secondly, it is a widely known and understood algorithm, having been implemented and tested rigorously for over a decade in a practical situation. Since it is so widely known and understood, and since its method of modeling page rank within a network as a flow through that network closely resembles the notion of influence flow which we would like to capture, it serves as a good first attempt at modeling the *influence* portion of our analysis. Although there are some subtle differences, the crux of the algorithm remains the same, however, and, for reasons that will become clear after the discussion, serves as a good initial attempt at modeling our concept of influence quantification. As the bulk of the algorithm is so widely known and understood, we don't spend much time here on the common details, as they can be found in [1]. Rather, we discuss the points here which differ from the original algorithm.

This said, we define an iterative method for computing a quantification of influence within a blogspace network, modeled on that of [1], whereby we assign an initial distribution of influence throughout the network and refine these values through a series of iterations

$$I_1(B,\mathcal{T}), I_2(B,\mathcal{T}), ..., I_n(B,\mathcal{T})$$

until for all B, $I_n(B,\mathcal{T}) - I_{n-1}(B,\mathcal{T}) < \varepsilon$ for some small ε (i.e., until the values converge).

Whereas the algorithm given in [1] assumes a uniform initial probability distribution, our initial distribution is nonuniform, reflecting instead the factors given by criteria 1–3 (*number of followers, relevancy, comments*) above:

$$I_0(B,\mathcal{T}) = |\text{IN}(B)| \cdot \pi(B,\mathcal{T}) \cdot \tau(B,\mathcal{T}) \quad (14)$$

Observe that this initial distribution imparts several desirable traits. First, if B has zero followers, the distribution for B falls to zero, reflecting the notion discussed above that regardless of what is written, a blog cannot be considered to be influential if it has no followers. Conversely, if B has many followers, this factor cases the distribution to increase accordingly. Secondly, if blog B contains zero posts on topic $\mathcal{T}$, then this factor, and consequently the distribution, falls to zero, reflecting the notion that regardless of the number of posts on any other topic, a blog cannot be considered to be influential on $\mathcal{T}$ if there are no posts on this topic. Conversely, if all posts in B deal with topic $\mathcal{T}$, this factor becomes 1, giving full weight to the remaining factors. Similarly, the number of comments on relevant posts is reflected in the factor $\tau(B,\mathcal{T})$. By definition, this factor considers only posts in B dealing with topic $\mathcal{T}$. At one extreme, if all posts on this topic have a sufficient number of comments (i.e., above τ_{max}), then this factor evaluates to 1, giving full weight to the other factors. At the other extreme, if all posts on this topic have zero comments, then this factor evaluates to 0, reflecting the notion that the blog cannot be considered to be influential on a topic if no comments are ever made on any posts dealing with that topic.

The subsequent iterations of the algorithm remains the same:

$$I_{i+1}(B,\mathcal{T}) = (1-\delta) + \delta \cdot \sum_{C \in \text{IN}(B)} \frac{I_i(C,\mathcal{T})}{|\text{OUT}(C)|} \quad (15)$$

As in [1], the *damping factor* δ serves the purpose of adjusting each successive influence value downwards, eventually ensuring convergence. The first term $(1-\delta)$ models the assumption, as in [1], that all blogs in the blogspace link to all other blogs (the so-called *random jump factor*). The second term— the summation of the influence quantities of each incoming link— quantifies exactly the notion that if a blog B which posts heavily on topic $\mathcal{T}$ is itself followed by a blogger who is influential in that same topic, then this should speak favorably on B's measure of influence on that topic as well. The optimal value of δ for our purposes is unknown, and will be computed empirically as part of future work. As mentioned earlier, Equation 15 is exactly the methodology given by [1], and as that work is now well-known and is not the focus of this paper, we leave the details to [1] instead.

We conclude this section by noting that our proposed method of computing influence has given us an algorithm which adheres to the criteria set forth at the beginning of this section. Namely, criteria 1–3 (number of followers, relevancy, and comments) are accounted for in the initial influence distribution across all blogs, and criteria 4 (followers) is accounted for by iterating Equation 15 until convergence, c.f. [1]. Finally, combining the sentiment analysis of Section 3.2 with the influence analysis of this section, gives us a concrete method of computing the endgoal for our analysis, namely the *influence-weighted sentiment* of a blogspace, with respect to a topic $\mathcal{T}$:

$$\sigma_I(\bar{B},\mathcal{T}) = \sum_{B \in \bar{B}} \sigma(B,\mathcal{T}) \cdot I(B,\mathcal{T}) \quad (16)$$

Notably, compare the notion of weighted sentiment given by Equation 8, where each blog is weighted only by the proportion of posts on a topic, with this one, in which each blog is weighted by its measure of influence in that topic.

4. DISCUSSION

Our model raises several issues which merit some discussion. Some of these issues are easily addressed, and some are less so.

First, one might initially consider the fact that our algorithms are designed with respect to a precomputed topic hierarchy to be a weakness, when instead it might be more desirable to be able to compute the hierarchy concurrently. Computationally, however, these two models are equivalent— in the worst case, one could take two separate passes through the blogspace, computing the hierarchy on the first pass and using this precomputed hierarchy as input to the algorithms presented here on the second.

On a more concrete level, a practical implementation will have several technical issues to address as well. Our notion of *influence* takes into account the number of comments left on posts, with the idea that posts with more comments are deemed to be generating a higher level of audience feedback. Technically, however, this may not always be the case. For instance, in a framework in which authors have the option of disabling reader comments, a post may have zero comments simply because comments have been disabled, not necessarily indicating a lack of audience interest. In this case, while clearly not an optimal situation, one possible compromise might be to use a heuristic— for example, assigning $|\textsc{Comm}(p)|$ a value of $\frac{\tau_{max}}{2}$ when this is the case. Similarly, a blogger may choose not to make public his list of followers and/or those he follows, in which case we have little choice but to assume that these lists are empty.

On a more subtle note, recall that our analysis computes a post's sentiment as a measure of the sentiment of the post's comments, albeit *with respect to the post's text*. While seemingly straightforward, this latter point raises a subtle yet strikingly important detail. Namely, a comment of positive sentiment may in fact denote an overall negative sentiment towards the respective topic, if the post's text was negative as well. For example, suppose the topic of the post is a piece of legislation newly passed by the government of country X. Further, suppose that all comments on this post are deemed to be overwhelmingly positive. One might initially be tempted to conclude that readers of this post are overwhelmingly in favor of the newly passed legislation. However, if the post itself speaks negatively about this legislation, the conclusion should be exactly the opposite; in this case, the overall audience sentiment is strongly negative. Thus there is a delicate interaction between the sentiment of the comments and that of the post itself which, in the interest of simplifying the presentation, we have chosen to omit from our analysis. However, this is a point that, while not exceedingly difficult, our implementation will need to address.

Finally, we observe that our sentiment score for a topic $\mathcal{T}$ gives equal weighting amongst all subtopics of $\mathcal{T}$, an assumption that is perhaps overly simplistic. Further refinement of this analysis, assigning varying weights among subtopics, remains an area of significant interest.

5. FUTURE WORK AND CONCLUSION

In this paper, we have defined a formal framework to model blogs, their posts and comments, sentiment, and influence. In the context of gauging public sentiment with respect to a given topic hierarchy, we give algorithms for computing influence and sentiment, which can then be used to give a measure of sentiment weighted by each blog's level of influence. As blogs by their very nature are constantly evolving, we prove that our formalization lends itself to efficient updates. We proved this by giving upper time bounds in the form of algorithms for processing such updates.

While we have presented our theoretical foundations in this paper, there is much left to be done. For instance, regarding sentiment, the questions raised in Section 4 regarding the potential ambiguity between sentiment on an *issue* versus sentiment on the post itself needs to be addressed, and doing so will not be a trivial task. Regarding influence, an obvious question is to determine the correct value for δ; this was computed empircally in [1], and we envision a similar approach in our context. Finally, and perhaps most importantly, we intend to implement the concepts presented here as a framework into which blogs can be harvested and analyzed using the techniques presented here.

6. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.

[2] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L. Tseng. On evolutionary spectral clustering. *ACM Trans. Knowl. Discov. Data*, 3(4):1–30, 2009.

[3] Yun Chi, Belle L. Tseng, and Junichi Tatemura. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 68–77, New York, NY, USA, 2006. ACM.

[4] Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, and Belle L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 163–172, New York, NY, USA, 2007. ACM.

[5] Yoonjung Choi, Youngho Kim, and Sung-Hyon Myaeng. Domain-specific sentiment analysis using contextual feature generation. In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 37–44, New York, NY, USA, 2009. ACM.

[6] Scott Farrar and Steve Moran. The e-linguistics toolkit. In *e-Humanities—an Emerging Discipline: Workshop in the 4th IEEE International Conference on e-Science*. IEEE, 2008.

[7] Jonathan G. Fiscus and George R. Doddington. Topic detection and tracking evaluation overview. pages 17–31, 2002.

[8] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

[9] Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM.

[10] Ben He, Craig Macdonald, Jiyin He, and Iadh Ounis.

An effective statistical approach to blog post opinion retrieval. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1063–1072, New York, NY, USA, 2008. ACM.

[11] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[12] Pyung Kim and Sung Hyon Myaeng. Usefulness of temporal information automatically extracted from news articles for topic tracking. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(4):227–242, 2004.

[13] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[14] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 375–384, New York, NY, USA, 2009. ACM.

[15] Yu-Ru Lin, Yun Chi, Shenghuo Zhu, Hari Sundaram, and Belle L. Tseng. Analyzing communities and their evolutions in dynamic social networks. *ACM Trans. Knowl. Discov. Data*, 3(2):1–31, 2009.

[16] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, and Belle L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 1–8, New York, NY, USA, 2007. ACM.

[17] Craig Macdonald and Iadh Ounis. Key blog distillation: ranking aggregates. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1043–1052, New York, NY, USA, 2008. ACM.

[18] Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Simple semantics in topic detection and tracking. *Inf. Retr.*, 7(3-4):347–368, 2004.

[19] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 533–542, New York, NY, USA, 2006. ACM.

[20] Prem Melville, Wojciech Gryc, and Richard D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284, New York, NY, USA, 2009. ACM.

[21] Lan Nie and Brian D. Davison. Separate and inequal: preserving heterogeneity in topical authority flows. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 443–450, New York, NY, USA, 2008. ACM.

[22] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, 2008.

[23] Arun Qamra, Belle Tseng, and Edward Y. Chang. Mining blog stories using community-based and temporal clustering. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 58–67, New York, NY, USA, 2006. ACM.

[24] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. Analyzing (social media) networks with nodexl. In *C&T '09: Proceedings of the Fourth International Conference on Communities and Technologies*, pages 255–264, New York, NY, USA, 2009. ACM.

[25] Xiaodan Song, Yun Chi, Koji Hino, and Belle Tseng. Identifying opinion leaders in the blogosphere. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 971–974, New York, NY, USA, 2007. ACM.

[26] Xiaodan Song, Ching-Yung Lin, Belle L. Tseng, and Ming-Ting Sun. Modeling and predicting personal information dissemination behavior. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 479–488, New York, NY, USA, 2005. ACM.

[27] Jintao Tang, Ting Wang, Ji Wang, and Dengping Wei. Efficient social network approximate analysis on blogosphere based on network structure characteristics. In *SNA-KDD '09: Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, pages 1–8, New York, NY, USA, 2009. ACM.

[28] Tun Thura Thet, Jin-Cheon Na, Christopher S.G. Khoo, and Subbaraj Shakthikumar. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 81–84, New York, NY, USA, 2009. ACM.

[29] Belle L. Tseng. Blog analysis and mining technologies to summarize the wisdom of crowds. In *MDM '07: Proceedings of the 8th international workshop on Multimedia data mining*, pages 1–1, New York, NY, USA, 2007. ACM.

[30] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, New York, NY, USA, 2006. ACM.

[31] Yiming Yang, Jaime Carbonell, Ralf Brown, John Lafferty, Thomas Pierce, and Thomas Ault. Multi-strategy learning for topic detection and tracking: a joint report of cmu approaches to multilingual tdt. pages 85–114, 2002.