Early Online Identification of Attention Gathering Items in Social Media

Michael Mathioudakis
Computer Science
University of Toronto
mathiou@cs.toronto.edu

Nick Koudas
Computer Science
University of Toronto
koudas@cs.toronto.edu

Peter Marbach
Computer Science
University of Toronto
marbach@cs.toronto.edu

ABSTRACT

Activity in social media such as blogs, micro-blogs, social networks, etc is manifested via interaction that involves text, images, links and other information items. Naturally, some items attract more attention than others, expressed with large volumes of linking, commenting or tagging activity, to name a few examples. Moreover, high attention can be indicative of emerging events, breaking news or generally indicate information items of interest to a vast set of people. The numbers associated with digital social activity are astonishing: in excess of millions of blog posts, tweets and forums updates per day, millions of tags in photos, news articles or blogs. Being able to identify information items that gather much attention in such a real time information collective is a challenging task.

In this paper, we consider the problem of early online identification of items that gather a lot of attention in social media. We model social media activity using *ISIS*, a stochastic model for *Interacting Streaming Information Sources*, that intuitively captures the concept of attention gathering information items. Given the challenge of the information overload characterizing digital social activity, we present sequential statistical tests that enable early identification of attention gathering items. This effectively reduces the set of items one has to monitor in real time in order to identify pieces of information attracting a lot of attention.

Experiments on real data demonstrate the utility of our model, as well as the efficiency and effectiveness of the proposed sequential statistical tests.

Categories and Subject Descriptors

J.7 [**Computer Applications**]: COMPUTERS IN OTHER SYSTEMS—*Real time*; H.m [**Information Systems**]: MISCELLANEOUS

General Terms

Measurement

Keywords

Social media analysis, User activity modeling and exploitation

WSDM'10, February 4–6, 2010, New York City, New York, USA.
Copyright 2010 ACM 978-1-60558-889-6/10/02 ...$10.00.

1. INTRODUCTION

Activity in social media such as blogs and micro-blogs (hosted by e.g., Blogger, Wordpress, LiveSpace, Twitter, Jaiku), social networks (e.g., Facebook, MySpace, Friendster), multimedia sharing services (e.g. Youtube, Flickr) or online newspapers and magazines has been increasing at a phenomenal pace. Millions of individuals participate daily in a social process of information exchange, generating information items such as blog posts, images, videos or status messages, as well as engaging with each other's generated items, e.g. by leaving comments or sharing them with friends. Indicative of the participation in social media are the 300 million users of MySpace and Facebook [1, 9], the more than 30 million regularly updated blogs [2], millions of users of Twitter, Youtube, Flickr, etc.

At an abstract level, individuals participating in social media can be thought of as **information sources** that *emit units of information* in a *streaming* fashion. Digital items such as blog posts, videos, pictures and short 'status' messages are all examples of information units. Besides acting as information sources, individuals also **interact** with each other. For instance, friends in a social network such as Facebook or Friendster visit each other's profiles to view the newly updated status messages or posted pictures and possibly engage with them. Engaging with an item involves performing **actions** such as leaving a comment, rating it or recommending it to others who might find it interesting.

Naturally, some generated items gather more attention than others. For example, blog posts, pictures or videos related to important emerging events often attract significant number of links and comments in a few hours. Distinguishing those items among the plethora of items generated in social media necessitates the definition of a measure for **attention gathering potential**, i.e. the 'ability' of items to attract their audience's attention and stimulate their reactions. In the case of blogs, for example, common measures include the total number of attracted links or comments, the number of distinct linkers (as it is the case with Technorati [3]), etc.

Such measures, however, fail to capture significant temporal aspects of social media activity. For instance, consider a blog post p_1 that attracts 10 links after remaining on the front page of its hosting blog for 1 week. Consider, as well, a blog post p_2 that also attracts 10 links, but only after remaining on the front page of its hosting blog for 1 hour. Taking into account the time each post remained visible on a blog webpage, it is reasonable to claim that post p_2 is associated with higher potential in attracting links than post p_1, even though the total number of links is the same. As another example, consider different blogs that are visited with varying rates by their readers. A post p_3 published on a blog that is rarely visited by its readers is less probable to attract the same number of actions (links or comments) with a post p_4 published on a frequently visited blog. Therefore, in case p_3 attracts the same number of actions with p_4, that fact should be interpreted as p_3 having larger attention gathering potential than p_4. These examples indicate that it is more intuitive to measure the attention gathering potential of items by taking into account not only the total number of actions they attract, but also temporal aspects of social media activity.

To capture such temporal aspects, we propose a novel measure of attention gathering potential that encompasses the temporal dimension. The measure is derived from the analysis of **ISIS**, a general stochastic model for *Interacting Streaming Information Sources* that is carefully defined to intuitively follow the way individuals in social media generate and engage with each other's items. In what follows, items with large attention gathering potential will be referred to as 'attention gathering items'.

Activity in social media is a dynamic process, with a large number of new items generated continuously and attention constantly shifting among items. In such a dynamic setting, it is important that attention gathering items are identified in real time, as social media activity evolves. For example, if a recently published blog post reports an interesting story that attracts a significant number of links from other sites and comments from its viewers, it is preferable to report it in real time, as it might correspond to important emerging news (a crisis, accident, announcement, etc). Therefore, identification of attention gathering items is best suited as an online – rather than offline – task. Also, given the large volume of data such a task needs to process in real time, it is more efficient to *prune* from consideration as early as possible items that do not appear likely to attract much attention and focus on monitoring a smaller candidate set of items with larger attention gathering potential.

A heuristic way to identify attention gathering items in online fashion is the following: "Maintain the number of actions each item attracts over time and report as 'attention gathering' the ones that exceed a threshold k. Also, discard items that do not exceed the threshold after dt time from their creation." However, setting the parameters k and dt in a meaningful way that takes into account temporal aspects of social media activity, is a non-trivial issue: How would k be set for sources that interact at different rates with other sources? If we wish to prune items that do not gather attention, what would be the 'right' value for dt so that we discard them early, but also avoid missing items that exceed threshold k later?

To address the aforementioned issues, we present a principled approach that uses sequential statistical tests in order to achieve early online identification of attention gathering items. The tests are based on the assumptions of the ISIS model and allow for the exploration of trade-offs between early reporting of results and quality. Experiments over real data from social media activity demonstrate that this approach can achieve significantly early identification of attention gathering items, compromising little quality in its results.

To summarize, we make the following contributions:

- We propose and analyze ISIS, a general stochastic model for interacting streaming information sources.
- Under ISIS, we derive a measure for the attention gathering potential of information units, that incorporates temporal aspects of social media activity in an intuitive way.
- We present sequential statistical tests for early online identification of items with large attention gathering potential.
- We present experimental results on real data collected from a period of blogging activity. The experiments demonstrate the application of the model in real-world scenarios and attest to the efficiency and effectiveness of the proposed statistical tests for early identification of attention gathering items.

To the best of our knowledge, this is the first work in the context of social media that formalizes and addresses the problem of *early online identification of attention gathering items*.

1.1 Roadmap

The paper is organized as follows. Connection with previous work is discussed in section 2. The technical part of the paper is covered by sections 3 and 4. In section 3, we describe ISIS, a model that intuitively follows the way social media activity evolves and we propose a measure for attention gathering potential. Subsequently, based on the model and its analysis, section 4 describes how sequential tests are used in order to achieve early, online identification of items with large attention gathering potential. Section 5 provides experimental results from the analysis of a blogging activity period that demonstrate the trade-offs in the performance of the sequential tests over real data. The paper concludes with section 6.

2. RELATED WORK

Link analysis has been widely used to obtain measures for the 'importance' of webpages [7, 8, 17, 20]. Conventionally, webpages are modeled as nodes of a graph, with directed edges between nodes corresponding to hyperlinks between webpages [18]. Importance values are then obtained for each webpage using graph-related measures – for example, the PageRank of a webpage is such a graph-based measure ([7] provides an in-depth summary of link analysis approaches). Yet, the way social media activity evolves suggests a departure from the traditional web model. For instance, linking in social media is explicitly associated with individual documents, pictures, news articles, etc and not just with the webpages that host those items. Therefore, it is reasonable to have separate measures for the importance or *attention gathering potential* of different items. Moreover, linking activity in social media is the product of continuous interaction between participating individuals. Dynamic aspects of this process (such as the *rate* with which content is generated or interactions occur) are not captured by the graph model, since it only considers the total number of links between webpages. Finally, linking is not the only action by which structure arises in social media, as individuals also interact by commenting, sharing, recommending or rating items they encounter online. In summary, in the case of social media, individual webpages are better modeled as information sources that emit information units in a streaming fashion and interact by dynamically performing different types of actions upon each other's units. In this work, we provide a first formal definition and analysis of such a model and use it as a basis to identify attention gathering items in online fashion.

Since attention gathering items possibly point to emerging events, our work has some affinity to event detection [4, 10, 11, 16, 19, 23]. Note, however, that there are strong dissimilarities between the two. As described in [4], the goal of event detection is to identify stories over a collection or stream of documents. Text analysis is applied towards that end, possibly taking into account linking activity or an underlying social network structure [23]. On the contrary, our work identifies individual items that attract a significant number of actions and its main focus is 'early identification' of such items – i.e. given a definition of what constitutes an 'attention gathering item', identify it as early as possible.

3. THE ISIS MODEL

In this section, we present **ISIS**, a model for interacting streaming information sources, that intuitively follows the way social media activity evolves and captures the concept of attention gathering information items. The formal definition of the model is given in section 3.1 and its analysis is provided in section 3.2. Notation used in this section and throughout the paper is summarized in table 1.

T	time period under study
u	streaming information source
U	the set of sources
p	information unit
P_u	set of units emitted by source $u \in U$
P	set of all units generated during T
t_p	creation time of unit p
d_p	validity period of unit p
λ_u	the rate at which other sources interact with source u
Λ	the set of interaction rates for all sources in U
x	action
t_x	timestamp of action x
w_p	interaction weight of unit p
W	set of interaction weights of all units in P
X_p	the set of actions attracted by unit p
X_u	the set of actions attracted by units of source u
X	the set of all actions attracted by units in P

Table 1: Notation

3.1 Model Definition

The purpose of the ISIS model is to serve as an abstraction of social media activity. Information sources (or 'sources', for simplicity) in the model correspond to individuals contributing information. A source is assumed to participate in two sets of stochastic processes:

1. The process of **emitting** information units in a streaming fashion.

2. Processes of **interaction** with other sources.

Information units (or, simply, 'units') emitted by sources conceptually correspond to items such as blog posts, status messages, photos, etc that appear in the social media stream. Interaction between sources corresponds to individuals engaging with each other's items. Thus, *interaction between sources* is assumed to involve sources performing *actions* upon units emitted by other sources. In the ISIS model, the notion of 'action' is used to represent different forms of engagement with items (such as linking, commenting, recommending, etc). The two sets of processes are subsequently described in detail (Subsections 3.1.1 and 3.1.2).

3.1.1 Emission of Information Units

Consider a set U of streaming information sources. A source $u \in U$ **emits** units according to a stochastic process, with every arrival of the process corresponding to the emission of a unit p (Figure 1). For example, units emitted by a source might correspond to blog posts published on a blog. Each unit p is associated with two time values, a timestamp t_p and a validity period d_p, both of which are known (observed variables).

Timestamp t_p denotes the time when unit p is emitted. For example, posts published on a blog or status updates on micro-blogging websites (e.g. Twitter) are accompanied by a timestamp declaring the time the post or status update was generated.

Validity period d_p is used to model the temporary nature of social media activity, i.e. the fact that items generated in social media do not remain relevant, interesting or available to the public for an infinite amount of time. For example, readers of a blog are not expected to read and comment on posts that were created a long time ago or have been removed from the front page of that blog. In practice, we consider the validity period to be equal to the time interval for which a post, news article, status update or other item remains on the front page of the related blog, news portal, social network profile, etc. During the validity period of a unit, we refer to the unit as *valid*.

For the purposes of the analysis that follows in section 3.2, assume that all quantities refer to social media activity that takes place during a time period T. In particular, let P_u denote the set of all units p emitted by source $u \in U$ and P denote the set of all units.

$$P = \bigcup_{u \in U} P_u$$

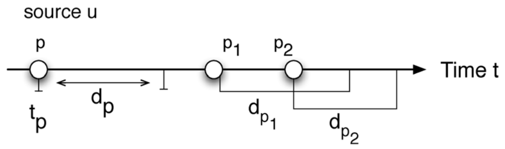


Figure 1: Information source. Each unit is associated with a timestamp t_p and a validity period d_p. Notice that validity periods of units emitted by the same source might overlap.

3.1.2 Interactions between Streaming Information Sources

Besides emitting information units, sources also **interact** with each other – e.g., friends in a social network interact by visiting each other's profile webpage. Moreover, during interactions of a source u' with another source u, there is a probability that u' performs an **action** upon valid units of u. For example, while individuals interact with their friends in a social network by visiting their profiles, they *sometimes* perform an action (e.g. leave a comment) upon their friends' posted items (pictures, status updates, etc). The time t_x an *action* x occurs is known (observed variable). For instance, when a person leaves a comment on an item, the comment is accompanied by a timestamp that declares the time the action took place.

In general, it is not possible to know when interactions occur, unless they involve an action. For example, it is not possible to determine when friends in a social network *visit* each other's profiles, as browsing history information is only available to the administrator of the social network website. Consequently, the time interval δt between successive interactions of source u' with another source u is a latent (unobserved) variable.

In interest of simplicity, the process by which interactions occur is assumed *memoryless*. Specifically, a source $u_i \in U$ is assumed to interact with source $u_j \in U \setminus \{u_i\}$ according to a Poisson process $\mathbb{I}_{u_i,u_j}(\lambda_{u_j})$ of rate λ_{u_j} [12] , with every arrival of the process corresponding to a single interaction of u_i with u_j. Equivalently, for any two successive interactions of u_i with u_j at times t_k and t_{k+1}, the inter-arrival interval $\delta t = t_{k+1} - t_k$ of process $\mathbb{I}_{u_i,u_j}(\lambda_{u_j})$ is the value of a random variable Δt that follows an exponential distribution with parameter λ_{u_j}.

$$Pr(\Delta t = \delta t) = Exp(\lambda_{u_j}) = \lambda_{u_j} e^{-\lambda_{u_j} \delta t}$$

Interaction rate λ_u is a latent variable, the value of which can be estimated based on the values of observed variables, as explained in detail in section 3.2.

Notice that interactions between sources are not assumed to be symmetric, i.e. the process according to which a source $u \in U$ interacts with another source $u' \in U$ is assumed distinct and independent from the process according to which u' interacts with u. Notice also that the rate λ_{u_j} with which source u_i interacts with source u_j is assumed to depend on u_j only and it will be referred to as the **interaction rate** of source u_j[1]. In what follows, Λ will be used to denote the interaction rates of all sources.

$$\Lambda = \{\lambda_u | u \in U\}$$

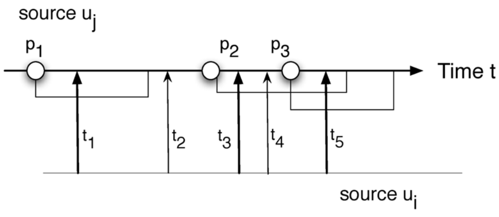


Figure 2: Source interaction.

During interaction of source u' with source u, u' might perform an action upon a valid unit p of u. More specifically, it is assumed that each valid unit p emitted by a source u is associated with an **interaction weight** w_p that determines the probability that a source u' which interacts with source u performs an action upon p. For example, when a blog post p is viewed by a reader for the first time, the reader leaves a comment to p with probability w_p.

Interaction weight w_p is not known a-priori (it is a latent variable). However, it can be estimated, given the number of actions unit p attracts, its validity period d_p and the interaction rate λ_u of source u. At a high level, the values of d_p and λ_u determine the number of interactions of sources u' with source u that occur while p is valid and therefore how many 'chances' unit p has to attract an action. The smaller the values of d_p and λ_u, the smaller is the expected number of such interactions. Therefore, for a given number of actions attracted by unit p, the smaller its validity period d_p and/or interaction rate λ_u, the larger the estimated value of w_p. On the other hand, for fixed values of d_p and λ_u, the larger the number of actions attracted by unit p, the larger its estimated w_p. This connection between interaction weight w_p and other variables (number of actions, d_p and λ_u) is shown analytically in section 3.2 and experimentally in section 5.2.

We propose and use the estimated value of w_p as a measure for the **attention gathering potential** of items. In contrast with measures based solely on the number of actions an item attracts, the estimated value of w_p is not only a function of the number of actions, but it also depends on temporal aspects of social media activity captured by d_p and λ_u and has an intuitive interpretation as a probability value in the ISIS model. Estimation of w_p through maximum likelihood analysis will be the subject of section 3.2.

In formal terms, if an arrival of process $\mathbb{I}_{u_i,u_j}(\lambda_{u_j})$ occurs at time t, source u_i performs an *action* x upon unit p emitted by source u_j with probability

$$Pr^{action}(p) = w_p$$

as long as unit p satisfies the following two constraints: (1) p is valid at time t and (2) t is the first time u_i interacts with u_j while unit p is valid. The two constraints are imposed to model in a simple manner the temporary nature of social media activity, i.e. the fact that items do not attract actions for an infinite amount of time. If any of these two constraints is not satisfied, u_i performs **no action** upon unit p. Each action x is associated with a timestamp t_x that denotes the time it occurred and which coincides with the time of the corresponding arrival of process $\mathbb{I}_{u_i,u_j}(\lambda_{u_j})$.

In the example of figure 2, each arrow corresponds to an arrival of process $\mathbb{I}_{u_i,u_j}(\lambda_{u_j})$ and thus to an interaction of u_i with units emitted by u_j. According to this specific example, interactions occur at times $t_1, t_2, \ldots, t_5$ and according to ISIS, source u_i might perform an action upon units p_1, p_2, p_3 at times t_1, t_3, t_5, respectively, each time with probability w_{p_i}, $i = 1, 2, 3$. However, it cannot perform an action upon any item at time t_2, since there is no valid unit emitted by source u_j at that time, nor at time t_4, since u_i had already interacted with u_j at time t_3, while unit p_2 was still valid.

In principle, one can model different types of actions with different w's associated with each of them (i.e. use different w's for the actions of linking, commenting and so on). In interest of simplicity, a single type of action is assumed in this work; however extension of the model to more than one types of actions is straightforward.

In what follows, W will be used to denote the set of interaction weights for all emitted units $p \in P$

$$W = \{w_p | \text{ emitted unit } p \in P\}.$$

In addition, let X_p denote the set of actions x along with their associated timestamp t_x attracted by a single unit p and X_u denote all actions (together with their timestamps) attracted by units p of source u. (Notation $p \in u$ will be used to denote that unit p has been emitted by source u). X will denote the entire set of actions created during T.

$$X_u = \bigcup_{p \in u} X_p, \quad X = \bigcup_{u \in U} X_u$$

3.2 Analysis

In this section, a maximum likelihood analysis for ISIS is provided. The purpose of the analysis is to estimate the values of latent variables W and Λ given the values of the observed variables.

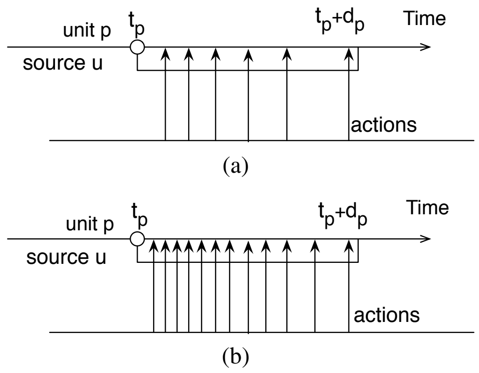
Consider the two examples shown in figure 3. Both depict source u emitting a unit p at time t_p, with the unit remaining valid for period d_p. However, in figure 3(a) unit p attracts a small number of actions in total, while in 3(b) unit p ends up attracting many actions. If an estimate has to be derived for the respective interaction weights w_0, w_1 of unit p for the two cases, then, since the number of actions attracted by unit p in fig. 3(b) is more than in fig. 3(a) in the same time period, interaction weight w_1 will be larger than w_0.

$$w_0 < w_1$$

In other words, since under ISIS a higher value of w_p implies a larger expected number of attracted actions $|X_p|$ for unit p, then maximum likelihood analysis returns higher estimates for w_p when a larger number of actions $|X_p|$ is observed.

However, the number of actions $|X_p|$ attracted by a unit p emitted by source u does not depend only on w_p. In fact, besides w_p, $|X_p|$ also depends on the validity interval d_p of unit p and the interaction rate λ_u of source u. Specifically, a larger validity interval d_p or interaction rate λ_u implies a larger expected number of actions $|X_p|$ under ISIS. Therefore, the same value of $|X_p|$ might

[1] One could consider a more general model with a distinct interaction rate λ_{u_i,u_j} for each pair of sources u_i, u_j. However, in order to keep the presentation and analysis of the ISIS model as simple as possible, we make the assumption that all sources u_i interact with u_j at the same rate λ_{u_j}.



(a)

(b)

Figure 3: Small and Large w_p.

lead to different (smaller or larger) estimates for w_p, depending on the value of d_p or λ_u. For example, if two units $p_1 \in u$, $p_2 \in u'$ have the same number of attracted actions

$$|X_{p_1}| = |X_{p_2}|$$

but different validity periods

$$d_{p_1} < d_{p_2}$$

then, unit p_1 will have a larger estimated interaction weight than p_2, since it attracted the same number of actions in less time.

$$w_{p_1} > w_{p_2}$$

3.2.1 Maximum Likelihood Estimation

Assume that the sets of observed variables P and X are available for a period T. Based on their values, the maximum likelihood values for the sets of latent variables Λ and W are computed.

Let $u \in U$ be a source that emits a sequence of units $P_u = [p_1, p_2, \ldots]$ during time period T and let $X_p = [x_1, x_2, \ldots]$ be the set of actions attracted by unit p. Then, the log-likelihood function of the latent variables is given by the formula

$$L(W, \Lambda) = \log Pr(X|P; W, \Lambda) = \sum_{u_j \in U} \sum_{p \in P_{u_j}} L_p(w_p, \lambda_{u_j}) \qquad (1)$$

with

$$L_p(w_p, \lambda_{u_j}) = \log Pr(X_p | t_p, d_p; w_p, \lambda_{u_j}), p \in u_j \qquad (2)$$

being the log-likelihood of a unit p attracting actions X_p, given its time values t_p, d_p, its interaction weight w_p and the rate λ_{u_j} of interactions with source u_j. Calculating the r.h.s. expression of equation 2 (details omitted due to space restrictions), we get

$$L_p(w_p, \lambda_{u_j}) = |X_p| \log(\lambda_{u_j} w_p) - \lambda_{u_j} \cdot \sum_{x \in X_p} (t_x - t_p) + (N - |X_p|) \log(1 - q_p w_p) \qquad (3)$$

with $N = |U|$ being the total number of sources participating and q_p being the likelihood that source u_i has interacted with u_j while unit p was valid.

$$q_p = 1 - e^{-\lambda_{u_j} d_p}$$

The first two terms of equation 3 represent the probability that a source u_i interacts with item p while it is valid *and* performs an action x at time t_x, while the third term expresses the probability this does not happen. Maximizing $L(W, \Lambda)$ requires W, Λ such that

$$\begin{cases} \nabla L(W, \Lambda) = 0 \\ 0 \leq w \in W \leq 1 \\ \lambda \in \Lambda \geq 0. \end{cases} \qquad (4)$$

System 4 can be solved using well known numerical methods [13, 21, 15]. A special case of the model that helps obtain better intuition upon the solutions of system 4 is analyzed in the following section.

3.2.2 A special case

According to the definition of ISIS (Section 3.1), units $p \in u$ are assumed to be associated with an interaction weight w_p, that determines the probability they attract an action from sources u' interacting with u. No further assumption is made about weights w_p, apart from the fact that they take values in the range $[0, 1]$. To gain some intuition into solutions of system 4, assume that all units p emitted by source $u \in U$ share the same interaction weight w_u

$$w_p = w_u, \quad p \in P_u \qquad (5)$$

and that the units are emitted and remain valid uniformly over time, i.e. that

$$d_p = d_u = k \cdot \frac{T}{|P_u|},\ p \in P_u,\ k = \text{ constant} \qquad (6)$$

where k denotes the number of items of source u that are valid at the same time. Assume also that all sources share the same interaction rate λ – which will be referred to as the *global interaction rate*.

$$\lambda_u = \lambda, \quad u \in U \qquad (7)$$

As it is easy to verify, for a *fixed* value of λ, the interaction weights $w_p = w_u$ of units p emitted by source u are given by the formula

$$w_p = w_u = \frac{|X_u|}{N \cdot |P_u| \cdot q_p} = \frac{|X_u|}{N \cdot |P_u| \cdot (1 - e^{-k\lambda_u\, T/|P_u|})} \qquad (8)$$

where $q_p = 1 - e^{-\lambda_u d_p} = 1 - e^{-\lambda_u d_u} = 1 - e^{-k\lambda\, T/|P_u|}$ is the probability a source u' interacts with source u while a particular unit p is valid. Let us consider two extreme cases for λ in relation with d_u.

Case 1: $\lambda k \frac{T}{|P_u|} = \lambda d_u \to 0$. Then,

$$q_p = 1 - e^{-\lambda d_u} \approx 1 - (1 - \lambda d_u) = \lambda d_u = k\lambda \cdot T/|P_u|$$

and

$$w_u = w_p \approx \frac{|X_u|}{N \cdot |P_u| \cdot \lambda \cdot d_u} = \frac{|X_u|}{N \cdot |P_u| \cdot \lambda \cdot k \cdot \frac{T}{|P_u|}} \propto$$

$$\propto \frac{|X_u|}{\lambda T} \propto |X_u|$$

Case 2: $\lambda k \frac{T}{|P_u|} = \lambda d_u \to \infty$. Then,

$$q_p = 1 - e^{-\lambda d_u} \approx 1 - 0 = 1$$

and

$$w_u = w_p \approx \frac{|X_u|}{N \cdot |P_u| \cdot 1} \propto \frac{|X_u|}{|P_u|}$$

The first case demonstrates that when the rate λ at which other sources interact with source u is *small* compared to the rate $|P_u|/T = d_u^{-1}$ at which source u emits units, then the estimated value of the interaction weight w_u is determined by the average number of actions *per interaction* $\frac{|X_u|}{\lambda T}$ and thus by the *total number* $|X_u|$ *of attracted actions*, since all sources u share a global interaction rate $\lambda_u = \lambda$.

On the other extreme, when rate λ is *large* in comparison with $|P_u|/T = d_u^{-1}$, then the value of w_u is determined by the average number of actions *per unit* emitted.

The value w_u, estimated under the assumptions of equations 5, 6 and 7, will be referred to as the **aggregate interaction weight** of source $u \in U$. As it is also the case with interaction weight w_p of individual items p, w_u has a simple and intuitive interpretation as a probability in ISIS and characterizes the *overall* 'ability' of a source u to attract actions from other sources with its units. Thus, just as interaction weights w_p are used to measure and compare the attention gathering potential of individual items, aggregate interaction weights w_u are used to perform a similar comparison at the source level.

4. EARLY ONLINE IDENTIFICATION OF ATTENTION GATHERING ITEMS

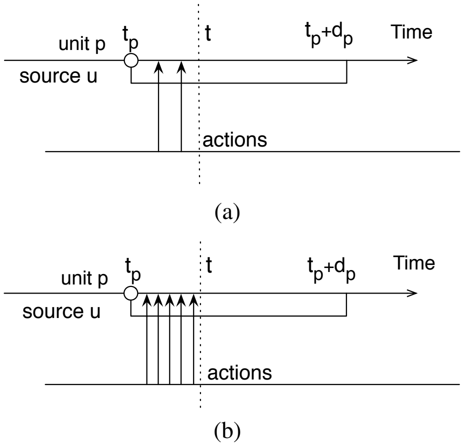
The ISIS model provides a formal framework for the estimation of interaction weights w_p of units p, as well as interaction rates λ_u of sources u, based on activity observed during a time period T. According to the analysis of section 3.2, estimation is achieved by solving system 4 and is performed in *offline* fashion, after data from period T is collected. In the context of social media monitoring, units p with large estimated interaction weight w_p correspond to items with high attention gathering potential.

In this section, we explain how ISIS is used when identification of items with high attention gathering potential needs to be performed in *online fashion* and return results *as early as possible*. The motivation for early, online identification of such items comes from the fact that they might contain information that refers to an evolving event (e.g. a crisis) or point to novel information that people deem important. When such items are identified, they are reported as 'attention gathering items'. At the same time, items that are not likely to be of large attention gathering potential are pruned from consideration. In this way, identification focuses on a smaller subset of candidate items.

For an illustrative introduction to the problem, consider the two cases depicted in figure 4. Similarly with the examples of figure 3, they involve source u emitting a unit p at time t_p that remains valid for a period of length d_p. However, unlike figure 3, figure 4 provides a 'snapshot' of the activity up to time t within the validity period of unit p.

Let w_1, w_0 be the interaction weights used in the examples of figures 3(b) and 3(a), respectively, with $w_0 < w_1$. The question addressed in this section is the following: having observed the interaction of sources $U \setminus \{u\}$ with source u only up to time t, is it possible to decide, with high confidence, whether unit p has a large interaction weight w_1 or a small interaction weight w_0? In fig. 4(b), for example, unit p has attracted many actions up to time t. Under the assumptions of ISIS, it is reasonable to predict that unit p will indeed end up with a large number of attracted actions until the end $(t_p + d_p)$ of validation period, just as in fig.3(b). Thus in the example of fig. 4(b), we have strong early indication that unit p has large interaction weight w_p rather than small and that it is more likely to be $w_p = w_1$ rather than $w_p = w_0$. On the contrary, the example of figure 4(a) is more similar to that of figure 3(a), and the small number of actions attracted by unit p up to time t indicate that its interaction weight w_p is more likely to be 'closer' to w_0 than to w_1.

More formally, consider a source u with known interaction rate λ_u. In practice, λ_u is estimated offline according to the analysis of section 3.2, based on activity of source u during a recent time period T. The value of λ_u is considered to remain relatively invariant over short time periods and we make sure that new estimates of its value are obtained regularly. For each unit p emitted by u, we wish to resolve as early as possible and with high confidence whether the interaction weight w_p associated with p is large or small, where 'small' and 'large' are quantified by two values of interaction weight $w_0^u < w_1^u$, that are given as input.



(a)

(b)

Figure 4: Early Identification.

Specifically, based on the actions X_p unit p gathers with time, we attempt to determine which of the two values, w_0^u or w_1^u is the more likely interaction weight of unit p. If w_1^u is decided to be the one, then unit p is reported, otherwise if w_0^u is the most likely value, it is ignored. In both cases, we require that a decision is taken with high confidence, i.e. that the probability our decision is mistaken is less than an error parameter ϵ.

The values of w_0^u, w_1^u can be specified in various ways. One option is that a user of a social media monitoring system who wishes to detect attention gathering items of source u in real time sets w_0^u and w_1^u, thus specifying what level of interaction weight constitutes an attention gathering item or not. A second option is to set them automatically. One way to do this is the following. If the interaction weight of units p generated by source u over a recent time period T have an average of $m_u = avg_{p \in u}(w_p)$ and a standard deviation of $s_u = std_{p \in u}(w_p)$, then w_0^u and w_1^u are set to

$$w_0^u = m_u \qquad w_1^u = m_u + 2 \cdot s_u.$$

The rationale for this selection of values is that if interaction weights w_p follow a Gaussian distribution with mean m_u and standard deviation s_u, then the probability of interaction weight higher than $m_u + 2s_u$ is less than 5% and thus it would be intuitive to report p as 'attention gathering'. Notice that, in this way, items of source u that are identified as 'attention gathering' are not necessarily ones with large interaction weight in absolute terms, but rather ones that have significantly large interaction weight relatively to the average estimate obtained from period T.

4.1 Early Identification

In this section, we explain how sequential tests [14, 22] offer a principled way to address the problem under discussion. For a source $u \in U$, two interaction weight values w_0^u, w_1^u are specified. For every unit p emitted by u, consider hypotheses H_0 and H_1.

$$H_0^p : w_p = w_0^u \qquad H_1^p : w_p = w_1^u \tag{9}$$

To decide which of the two hypotheses to accept, we apply sequential test S, which is summarized in table 2.

Specifically, we observe the actions X_p^t attracted by p in the time period $[t_p, t]$. Every time t such an observation is made, a decision about which hypothesis to accept is based on the *likelihood ratio*

$$r_t = \frac{L_1^t}{L_0^t} = \frac{Pr(X_p^t|t_p; w_1^u, \lambda_u)}{Pr(X_p^t|t_p; w_0^u, \lambda_u)},$$

where L_0^t, L_1^t are the likelihoods that unit p has interaction weight w_0^u or w_1^u, given that it attracts actions X_p^t till time t.

If r_t is sufficiently large, then hypothesis H_1 is accepted. How large r_t needs to be for H_1 to be accepted is specified by test S based on the error parameter ϵ. Similarly, if r_t is sufficiently small, then hypothesis H_0 is accepted. In the case that r_t is not small or large enough to decide which hypothesis to accept, the same procedure is repeated after a small time interval δs. In case the validity period d_p of unit p expires before a decision is made, a decision is immediately made in favor of the hypothesis that corresponds to the larger likelihood L_0^t or L_1^t at time $t = t_p + d_p$.

At any point in time, we maintain a set C of units that are candidates to be identified as attention gathering. All units are added to set C as soon as they are generated and remain its members for as long as test S has not decided which of the two hypotheses H_0^p or H_1^p to accept. Units are removed from set C when test S terminates. If hypothesis H_1^p is accepted, they are reported as 'attention gathering', otherwise they are discarded.

Table 2: Sequential Test S

	Condition	Decision
$t_p \le t \le t_p + d_p$	$r_t < \frac{\epsilon}{1-\epsilon}$	$w_p = w_0^u$
	$\frac{\epsilon}{1-\epsilon} \le r_t \le \frac{1-\epsilon}{\epsilon}$	no decision yet
	$\frac{1-\epsilon}{\epsilon} < r_t$	$w_p = w_1^u$
$t_p + d_p < t$	$L_1^t < L_0^t$	$w_p = w_0^u$
	$L_1^t \ge L_0^t$	$w_p = w_1^u$

Sequential test S allows for the exploration of a trade-off between error parameter ϵ (i.e. the probability a correct decision is reached before the test is truncated) and the number of observations the test collects before one of the two hypotheses is accepted. More specifically, the larger error ϵ, the smaller is the time needed for a decision to be reached. In section 5, this trade-off is displayed experimentally over real datasets and it is shown that early identification can be achieved by compromising little quality in the results.

5. EXPERIMENTS

We provide experimental results from the application of ISIS (Section 3) and usage of sequential tests [22] (Section 4) on real social media data. In particular, section 5.1 presents real examples of attention gathering items that prove that ISIS is able to identify items related to emerging events and/or of increased interest to social media audience. Subsequently, section 5.2 demonstrates the ability of interaction weight w to capture temporal aspects of social media activity. Finally, section 5.3 presents the trade-offs that arise from the usage of the sequential tests.

The dataset used in the experiments was collected from BlogScope [5, 6], a social media warehousing platform developed at the University of Toronto and which currently hosts a multi-terabyte collection of data from social media activity. In particular, experiments were performed over real data from a 15-day period of blogging activity (T = [May 1st 2008 - May 15th 2008]). The dataset consists of the activity of the 1000 most 'active' blogs in that period, i.e. the blogs that attracted the most links from their viewers. In total, the dataset contained $280k$ posts, as well as $180k$ links attracted from those posts.

The correspondence between blogging activity and ISIS is displayed in table 3. According to that, the set of blogs that were active during period T act as streaming information sources, with blog posts as their emitted units. Moreover, blogs interact when blog owners visit the webpage of other blogs and they perform the action of **linking** upon each other's posts – i.e. during interaction of blog u with blog u', blog u possibly creates a link towards a blog post p generated on blog u'. Given this correspondence, the notation used previously for the definition of ISIS (Section 3) will also be used for the description of the experiments.

Table 3: Correspondence between Model & Data

Blogging Activity	ISIS Model
Blog	Streaming Information Source
Post	Emitted Information Unit
Visit a Blog owner visits another Blog	Interaction between Sources
Link from a Blog to a Post	Action performed by a Source upon a Unit

5.1 Attention Gathering Items

In this part of experiments, we give examples of blog posts that are identified as attention gathering items under ISIS. Following section 4, for each post p generated by blog u we compare two hypotheses, H_0^p and H_1^p.

$$H_0^p : w_p = w_0^u \qquad H_1^p : w_p = w_1^u$$

If the average value of interaction weight w_p for posts p of blog u is $m_u = avg_{p \in u}(w_p)$ during a recent period of activity and standard deviation is $s_u = std_{p \in u}(w_p)$, then w_0^u and w_1^u are set as

$$w_0^u = m_u \qquad w_1^u = m_u + 2s_u.$$

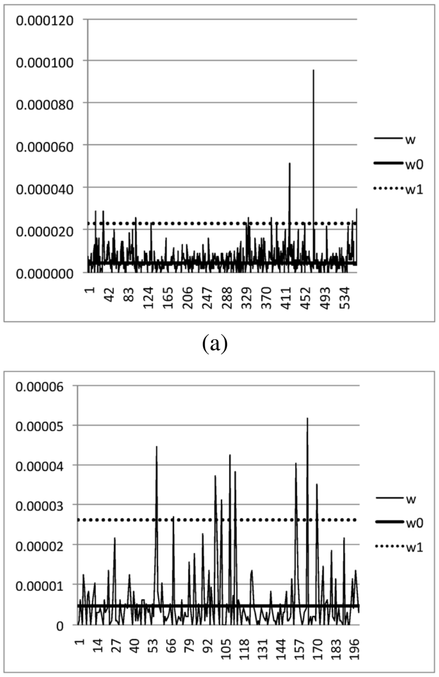
Test S is performed with error parameter $\epsilon = 0$ and posts p for which hypothesis H_1^p gets accepted are reported.

We present examples of posts that are reported as attention gathering items. The posts come from two specific blogs, i.e. engadget.com and techcrunch.com and the w_0^u,w_1^u values that were used for each blog during test S are mentioned in table 4. For illustration purposes, figure 5 contains the values of interaction weights w_p of posts belonging to the two blogs, as estimated by solving system 4. The bottom horizontal line in each plot inside figure 5 corresponds to w_0^u while the top horizontal line corresponds to w_1^u. The posts that were identified as 'attention gathering' from test S are the ones with interaction weight w_p that was closer to w_1^u than to w_0^u.

A sample of these posts (i.e. ones that were identified as attention gathering items) is shown in table 5. For example, on March 12th engadget.com published a post titled 'iPhone OS 3.0 is coming, preview on March 17th'[2]. The post reported on emerging news about the release of a new operating system for iPhone and attracted nearly 100 links from other blogs that also reported on the news and cited engadget.com as their source. Similarly, on March

[2] http://www.engadget.com/2009/03/12/iphone-os-3-0-is-coming-march-17th/

4th techcrunch.com published a post with title 'Facebook's response to Twitter'[3]. The post commented on the just announced change in Facebook's design and attracted a large number of links from other blogs that also commented on the news. The examples indicate that ISIS successfully identifies items related to emerging events or draw much attention from social media users.



(a)

(b)

Figure 5: Interaction weights of posts in (a) engadget.com (b) techcrunch.com.

Table 4: Blogs

Blog	w_0	w_1
engadget.com	$4.8 \cdot 10^{-6}$	$2.6 \cdot 10^{-5}$
techcrunch.com	$4.6 \cdot 10^{-6}$	$2.3 \cdot 10^{-5}$

5.2 Connection between Λ, W

In previous sections of this paper, we propose and use interaction weight w_p as an intuitive measure for the attention gathering potential of items p, that is not based only on the number of actions attracted, but also takes into account the temporal dimension of social media activity. The purpose of this part of experiments is to demonstrate that dependence of interaction weights on the temporal dimension through real examples. Towards that end, we apply the analysis of section 3.2 on the blogging activity dataset collected from BlogScope and exhibit the connection between latent variables Λ and W of ISIS.

In fact, in order to keep the presentation simple, we focus on the connection between *global interaction rate* λ and *aggregate interaction weight* w_u of blogs $u \in U$, defined in section 3.2.2. Following section 3.2.2, we assume that all blogs U share the same interaction rate λ

Table 5: Attention Gathering Posts

www.engadget.com
Microsoft shows a glimpse at the future of computing
Apple notebook in Q3
iPhone OS 3.0 is coming March 17th
Ubuntu 9.04 ported to Nokias N8x0 internet tablets'
Third party iPod Shuffle headphones will require license
Apple planning a March 24 event
www.techcrunch.com
Big music will surrender but not until at least 2011
GrandCentral to finally launch as Google Voice
Google privacy blunder shares your docs without permission
Wolfram Alpha computes answers to factual questions
Facebook's response to twitter
It's time to start thinking of Twitter as a search engine

$$\lambda = \lambda_u, u \in U$$

and that posts of the same blog $u \in U$ share the same length of validity period

$$d_p = d_u = k \cdot \frac{T}{|P_u|}, p \in u.$$

The value of constant k is set to $k = 50$, as it was experimentally found that for this value most (nearly 90%) of observed links were included in the validity periods of the corresponding posts. Based on these assumptions and solving system 4, we estimate the aggregate interaction weight w_u of blogs u based on three different and explicitly set values of λ ($\lambda = 10^{-5}h^{-1}, \lambda = 10^{-3}h^{-1}, \lambda = 10^{-1}h^{-1}$, $h = 1$ hour). For each value of λ, the 10 blogs with maximum w_u were computed and are reported in figures 6(a), 6(b) and 6(c), respectively.

The results in figure 6 are consistent with the theoretical analysis of section 3.2.2. More specifically, for small interaction rate λ, aggregate interaction weight w_u is determined by the *total number of links*. For this reason, the list of blogs with maximum w_u (Figure 6(a)) is dominated by the blogs with largest total number of links. On the other hand, for large λ values, w_u is determined by the *average number of links per post*. Consequently, the list of blogs with maximum w_u (Figure 6(c)) is dominated by blogs with the largest average number of links per post. Finally, when the value of λ is neither too big nor too small, the list of blogs with maximum w_u is mixed both with blogs with large total number of links or large average number of links.

The results demonstrate an intuitive relationship between *interaction weights* and *interaction rates*. In an setting where sources interact very frequently with each other, all generated units have the 'chance' to attract an action. Therefore it is reasonable to estimate the 'ability' of a source to attract actions by the average number of actions attracted *per unit*. On the other hand, in a setting where sources rarely interact with each other, the average number of actions attracted per unit is not a reasonable measure anymore: it would underestimate sources that emit many units, most of which do not actually get a 'chance' to attract an action while they are valid. In such settings, it is more intuitive to measure the 'ability' of a source to attract actions by the number of attracted actions *per interaction*, or simply the *total* number of actions if the interaction rate is equal for all sources. This argumentation can be extended from the source level to the level of units p and their interaction weight w_p. Under the described rationale, interaction weight w_p is an intuitive measure for attention gathering potential of items, that

[3] http://www.techcrunch.com/2009/03/04/facebooks-response-to-twitter/

Rank	Blog	Links	Posts	Links/Post
1	engadget.com	1812	563	3.2
2	guardian.co.uk.com	1324	1561	0.85
3	thinkprogress.org	1030	179	5.8
4	hotair.com	927	291	3.2
5	techcrunch.com	926	200	4.6
6	icanhascheezburger.com	862	93	9.3
7	michellemalkin.com	823	124	6.6
8	lifehacker.com	734	271	2.7
9	pajamasmedia.com	604	714	0.85
10	xkcd.com	519	7	74

(a) $\lambda = 10^{-5}\ h^{-1}$, $d_p = 50 \cdot \frac{T}{|P_u|}$

Rank	Blog	Links	Posts	Links/Post
1	engadget.com	1812	563	3.2
2	xkcd.com	519	7	74
3	guardian.co.uk	1324	1561	0.85
4	thinkprogress.org	1030	179	5.8
5	techcrunch.com	926	200	4.6
6	hotair.com	927	291	3.2
7	icanhascheezburger.com	862	93	9.3
8	michellemalkin.com	823	124	6.6
9	thestorybeginnings.blogspot.com	224	5	45
10	lifehacker.com	734	271	2.7

(b) $\lambda = 10^{-3}\ h^{-1}$, $d_p = 50 \cdot \frac{T}{|P_u|}$

Rank	Blog u	Links	Posts	Links/Post
1	xkcd.com	519	7	74
2	thestorybeginnings.blogspot.com	224	5	45
3	stevenberlinjohnson	107	3	36
4	grosgrainfabulous.blogspot.com	363	14	26
5	sethgodin.typepad.com	341	17	20
6	pinktentacle.com	116	6	19
7	asofterworld.com	156	9	17
8	funnyordie.com	449	26	17
9	smashingmagazine.com	330	20	17
10	lonelyheartscasino.com	128	8	16

(c) $\lambda = 10^{-1}\ h^{-1}$, $d_p = 50 \cdot \frac{T}{|P_u|}$

Figure 6: Blogs with maximum aggregate interaction weight for different λ values.

takes into account the temporal aspects of social media activity.

5.3 Quality vs Efficiency Trade-offs

The experiments presented in this section aim to demonstrate the performance benefits from utilizing the sequential tests described in section 4 as well as the arising trade-offs in efficiency and quality.

At a high level, quality in the experiments is measured by the fraction of correct decisions made by the sequential test w.r.t. the interaction weight of a post. Since for real data it is impossible to know the 'real' values of interaction weights, the classical measures of precision and recall cannot be used and we thus need to resort to measures that are based on experimentally defined 'ground truth'. The following seems to be a reasonable choice: *a post will be said to be 'true w_0' ('true w_1') when it is more likely to be of interaction weight w_0 (w_1) at the end of its validity period.*

Quality is measured through the following quantities: *experimental type I and type II error*, w_1 *impurity* and *true* w_1 *miss rate*. Experimental type I error expresses the fraction of true w_0 posts that are decided by test S to be of interaction weight w_1.

$$\text{type I error} = \frac{\#(\text{true } w_0 \wedge \text{decided } w_1)}{\#(\text{true } w_0)}$$

Similarly, experimental type II error expresses the fraction of true w_1 posts that are decided by the sequential test S to be of interaction weight w_0.

$$\text{type II error} = \frac{\#(\text{true } w_1 \wedge \text{decided } w_0)}{\#(\text{true } w_1)}$$

Moreover, w_1 impurity measures the fraction of posts identified as w_1 that are true w_0 posts.

$$\text{impurity} = \frac{\#(\text{true } w_0 \wedge \text{ decided } w_1)}{\#(\text{ decided } w_1)}$$

Finally, miss rate expresses the fraction of true w_1 posts that are either decided by test S to be of interaction weight w_0 or are not decided until their validity period expires.

$$\text{miss rate} =$$

$$= \frac{\#(\text{true } w_1 \wedge ((\text{decided } w_0) \vee (\text{test truncated})))}{\#(\text{true } w_1)}$$

$$= \text{type II error } + \frac{\#(\text{true } w_1 \wedge (\text{test truncated}))}{\#(\text{true } w_1)}$$

Efficiency is measured through average workload, i.e. the average number of posts over time that are considered as candidate attention gathering items. If $D_p \in [0, d_p]$ is the time it takes for a post to be decided either as attention gathering or be pruned from consideration, then

$$workload = \frac{\sum_p D_p}{T}$$

Similarly with the qualitative experiments for attention gathering items (Section 5.1), the hypotheses tested by sequential test S were

$$H_0^p : w_p = w_0^u = avg_{p \in u}(w_p)$$

$$H_1^p : w_p = w_1^u = avg_{p \in u}(w_p) + 2 \cdot std_{p \in u}(w_p)$$

with $avg_{p \in u}(w_p)$ being the average interaction weight w_p of posts p of blog u in a recent period of activity and $std_{p \in u}(w_p)$, being the standard deviation.

Figure 7 demonstrates the workload for different model errors ϵ. As shown in the figure, workload ranges from $50k$ posts when $\epsilon = 0$ and sequential test S is always truncated, to $5k$ posts when test S is performed with $\epsilon = 0.5$. Notice that by performing test S with $\epsilon = 0.05$, we already have a 50% decrease in the workload.

Figure 8 demonstrates the trade-off between the workload and the quality in its results, as measured by the experimental type I and type II errors (please note that workload is normalized w.r.t. its maximum value). As shown in the figure, if a type I error and type II error of 5% is tolerated, a 50% reduction in workload is achieved for $\epsilon = 0.05$.

Finally, it is interesting to notice that for small values of ϵ ($\epsilon < 0.275$) there is a trade-off between miss rate and impurity (Figure 9). Recall that miss rate corresponds to true w_1 posts that are either (a) identified as w_0 (type II error) or (b) posts that test S fails to identify before it is truncated. For small ϵ most missed true w_1 posts correspond to the second case. The interpretation of this trade-off is that, in order to make substantial use of the sequential

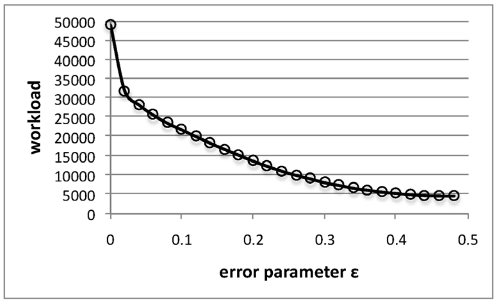


Figure 7: Workload vs Model Error

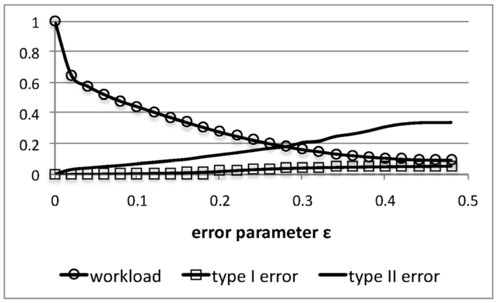


Figure 8: Trade-off between workload and type I, type II errors. Workload is normalized w.r.t. its maximum value.

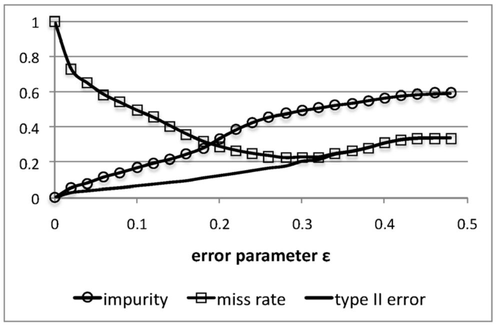


Figure 9: Trade-off between miss rate and impurity, for $\epsilon < 0.275$.

test (i.e. do not allow it to be truncated often), we need to accept some increase of the impurity of the identified attention gathering items (i.e. also identify some that are not true w_1). However, after some point ($\epsilon > 0.275$) the fraction of missed w_1 posts is dominated by the type II error and the trade-off between miss rate and impurity stops.

Overall, the experiments suggest that we can achieve significant decrease in the number of items we monitor by compromising little quality in the results.

6. CONCLUSIONS AND FUTURE WORK

This paper describes how sequential statistical tests are used to achieve early online identification of attention gathering items in social media. A significant part of our contribution lies with the definition and analysis of ISIS, a stochastic model for interacting streaming information sources that follows the way social media activity evolves. Its analysis leads to an intuitive measure for the attention gathering potential of items. Values of this measure are used in the sequential statistical tests as parameters that define which items are considered as 'attention gathering'.

In future work, we plan to explore possible extensions of ISIS to capture a wider range of behavior in the context of social media. In particular, we plan to explore the usage of stochastic modeling to capture early the development of viral phenomena or to monitor the formulation and dissolution of user communities.

Acknowledgements

We wish to thank Amalia Kokkinaki and Dimitris Tsirogiannis for their insightful comments on this paper.

7. REFERENCES

[1] Is facebook growing up too fast?, http://www.nytimes.com/2009/03/29/technology/internet/29face.html.

[2] State of the blogosphere 2008, http://technorati.com/blogging/state-of-the-blogosphere//.

[3] Technorati authority and rank, http://technorati.com/weblog/2007/05/354.html.

[4] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report, 1998.

[5] N. Bansal, F. Chiang, N. Koudas, and F. W. Tompa. Seeking stable clusters in the blogosphere. In *VLDB*, 2007.

[6] N. Bansal and N. Koudas. Blogscope: A system for online analysis of high volume text streams. In *WebDb*, 2007.

[7] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *TIS*, 2005.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 1998.

[9] Facebook. Latest facebook statistics, http://www.facebook.com/press/info.php?statistics.

[10] G. P. C. Fung, J. X. Yu, H. Liu, and P. S. Yu. Time-dependent event hierarchy construction. In *KDD*, 2007.

[11] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *VLDB*, 2005.

[12] R. Gallager. *Discrete Stochastic Processes*. 1st edition, 1995.

[13] T. Hoang. *Convex Analysis and Global Optimization*. 1998.

[14] P. Hoel. *Introduction to Mathematical Statistics*. 1984.

[15] R. Horst, P. Pardalos, and N. V. Thoai. *Introduction to Global Optimization*, volume 3 of *Nonconvex Optimization and Its Applications*. Springer, 1995.

[16] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD*, 2002.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.

[18] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models and methods. In *Lecture Notes in Computer Science*. 1999.

[19] A. Krause, J. Leskovec, and C. Guestrin. Data association for topic intensity tracking. In *ICML '06*, 2006.

[20] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. In *TIS*, 2000.

[21] J. D. Pintér. *Global Optimization in Action*. 1996.

[22] A. Wald. *Sequential Analysis*. Dover Phoenix Editions, 1947.

[23] Q. Zhao, P. Mitra, and B. Chen. Temporal and information flow based event detection from social text streams. In *AAAI*, 2007.