

# Eigen-Trend: Trend Analysis in the Blogosphere based on Singular Value Decompositions

Yun Chi  
NEC Laboratories America  
10080 N. Wolfe Rd, SW3-350  
Cupertino, CA 95014, USA  
ychi@sv.nec-labs.com

Belle L. Tseng  
NEC Laboratories America  
10080 N. Wolfe Rd, SW3-350  
Cupertino, CA 95014, USA  
belle@sv.nec-labs.com

Junichi Tatemura  
NEC Laboratories America  
10080 N. Wolfe Rd, SW3-350  
Cupertino, CA 95014, USA  
tatemura@sv.nec-labs.com

## ABSTRACT

The blogosphere—the totality of blog-related Web sites—has become a great source of trend analysis in areas such as product survey, customer relationship, and marketing. Existing approaches are based on simple counts, such as the number of entries or the number of links. In this paper, we introduce a novel concept, coined *eigen-trend*, to represent the temporal trend in a group of blogs with common interests and propose two new techniques for extracting eigen-trends in blogs. First, we propose a trend analysis technique based on the singular value decomposition. Extracted eigen-trends provide new insights into multiple trends on the same keyword. Second, we propose another trend analysis technique based on a higher-order singular value decomposition. This analyzes the blogosphere as a dynamic graph structure and extracts eigen-trends that reflect the structural changes of the blogosphere over time. Experimental studies based on synthetic data sets and a real blog data set show that our new techniques can reveal a lot of interesting trend information and insights in the blogosphere that are not obtainable from traditional count-based methods.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information filtering*; H.2.8 [Database Management]: Database Applications—*Data mining*; J.4 [Computer Applications]: Social and Behavioral Sciences—*Economics*

## General Terms

Measurement, Experimentation

## Keywords

Trend Analysis, Blog, Blogosphere, Singular Value Decomposition, Higher-Order Singular Value Decomposition

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.  
Copyright 2006 ACM 1-59593-433-2/06/0011 ...\$5.00.

## 1. INTRODUCTION

### 1.1 Motivation

A weblog, or *blog* in short, is a relatively new self-publishing phenomenon on the Web that is quickly becoming mainstream over the past few years. A blog is a special Web site on which an individual author (a *blogger*) or a group of collaborating authors periodically publish articles (*entries* or *posts*). Usually the entries are posted in reverse chronological order and each entry is attached with a time stamp indicating the time when the entry is posted. The world of blogs, which is commonly referred to as the *blogosphere*, is growing extremely rapidly—at the time of this writing, Technorati, one of the top blog search engines, is tracking more than 41.8 millions blogs. According to Technorati, more than 1.2 million new entries are created everyday. In addition, these numbers have been doubling every six months in the past three years<sup>1</sup>.

The blogosphere, an arena in which tens of millions of users share the latest information and exchange personal opinions, offers great commercial values and provides new business opportunities in areas such as product survey, customer relationship, and marketing. For example, in order for businesses to make judicious decisions, it is very important for them to track customer opinions and complaints in a timely fashion. Here the blogosphere provides free large-scale information sources from which businesses can quickly learn opinions and complaints from their customers. At the same time, as a special part of the Web, the blogosphere has its unique nature and features and therefore raises many new challenges. One such unique feature is that the blogosphere is much more dynamic than traditional Web pages. For example, an announcement of a new product may instantly trigger intensive discussions in the blogosphere. Very often, it is exactly these dynamic trends that are valuable for businesses to track, understand, and predict the interests of their customers. In this paper, we focus our study on trend extraction and analysis in the blogosphere.

### 1.2 Problems with Existing Solutions

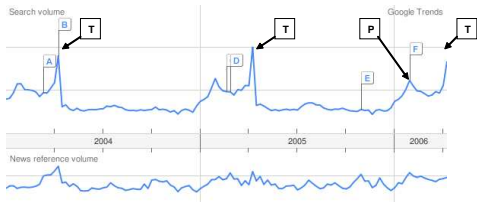
Recently, many commercial blog and Web search engines have introduced services for temporal trend analysis. For example, for given keywords, BlogPulse<sup>2</sup> and IceRocket<sup>3</sup> generate trend curves over time in terms of the percentage of

<sup>1</sup><http://technorati.com/weblog/2006/04/96.html>

<sup>2</sup><http://www.blogpulse.com/>

<sup>3</sup><http://trend.icerocket.com/>

blog entries that contain the keywords. For a given tag, Technorati<sup>4</sup> provides curves that show the daily number of entries that adopt the tag. Google has just announced a new service called Google Trend<sup>5</sup> that, for given keywords, plots the search volume and news reference volume that are related to the keywords over time. Figure 1 shows the curves given by Google Trend on the keyword *tax*. We use letter *T* to indicate the tax seasons (the middle of April). However, in the plot there is a peak, indicated by *P*, that happened at the beginning of February, 2006, which was not a tax season. The related news article listed by Google Trend further suggests that this peak was due to a political event that Israel halted tax payments to Palestinians.



**Figure 1: Google Trend Curve for Keyword *Tax***

All the commercial products mentioned above use time series of some simple statistics, such as percentage or total count, to represent temporal trends on the given keywords. Statistics such as total count or average have some good statistical properties and they usually describe central tendencies accurately. However, these statistics are aggregations and by using only these statistics, we lose some important insights into the blogosphere. More specifically, we raise the following questions and propose our solutions in this paper:

- (1) By summing up the occurrence of entries, traditional methods ignore individual blogs that published those entries. However, different blogs contribute to the trend differently. For example, some blogs constantly discuss on products by a specific company whereas others mention the company name occasionally (e.g., only when it is acquired by another company). Such differences in activity are not factored in by traditional methods. Is there any better way to represent the temporal behavior of the blogosphere by considering such differences among blogs?
- (2) For the same keyword, different groups of blogs may have different interests. Sometimes, a single trend does not make sense to all the interested groups of blogs. In the example shown in Figure 1, there are interest on tax from the financial point of view and interest from the political point of view. For a tax software company, the trend curve in Figure 1, which is an accumulation of all the interests, will be misleading for purposes such as supporting marketing decisions. On the other hand, in the blogosphere a blog usually does not explicitly indicate its interests (e.g., finance vs. politics). Can we automatically detect different groups of blogs with different interests and extract meaningful trends for the corresponding groups?

<sup>4</sup><http://www.technorati.com/tags/>

<sup>5</sup><http://www.google.com/trends>

- (3) The blogosphere is an ecosystem in which blogs interact with each other generating reference structure. In this sense, the blogosphere can be considered as a blog graph where the nodes are blogs and the links reflect endorsements and interactions among blogs. In addition, such a blog graph is changing with time as a result of the development of internal relationships (e.g., interactions among blogs) and external events (e.g., breaking news). Can we directly study and extract meaningful trends from such a dynamically changing graph structure?

### 1.3 Our Contributions

In this paper, we propose *eigen-trends*, temporal indicators derived through singular value decomposition, that take differences among individual blogs in consideration. The key idea is to represent the observed data as a combination of information that captures temporal changes of the underlying data (i.e., eigen-trends) and information that captures the characteristics of individual bloggers (e.g., authority). We show that this combination statistically gives an optimal estimation of the observed data. We propose two types of eigen-trends: *scalar eigen-trends* and *structural eigen-trends*, which are our main contributions:

- (1) We have developed a method based on the singular value decomposition (SVD) to extract multiple *scalar eigen-trends*. The main scalar eigen-trend best approximates the observed data and has good statistical properties. The secondary scalar eigen-trends can be used to represent non-dominating interests in the blogosphere.
- (2) We have developed a method based on a higher-order singular value decomposition (HOSVD) to extract *structural eigen-trends*. The structural eigen-trend detects the structural changes in the blogosphere.

Although SVD has been used for time-series analysis in areas such as meteorology [10], social science [14], and computer networking [15], to the best of our knowledge, we are the first one to apply a higher-order singular value decomposition for trend analysis of graph structure data.

Experimental studies based on synthetic data sets and a real blog data set show that these new techniques can reveal a lot of interesting trend information and insights in the blogosphere, which are not obtainable from traditional count-based methods, and therefore suggest that these new techniques can supplement traditional methods for trend analysis.

The rest of this paper is organized as follows. In Section 2, we give background information and related work. We present our trend analysis technique based on the singular value decomposition in Section 3 and that based on a higher-order singular value decomposition in Section 4. In Section 5, we study the proposed new techniques by using both synthetic data sets as well as a real blog data set. Finally in Section 6, we give the conclusion and future research directions.

## 2. BACKGROUND AND RELATED WORK

### 2.1 Background on Blogs

A *blog* is a Web site hosted by an individual or a group of collaborating authors (in either case, we call the host

a *blogger*) that publishes a series of articles (*entries*) over time. On the main page of a blog, there are the most recent entries by the blogger. The content of an entry consists of text and possibly other media (such as pictures, audio and videos). Each entry contains a time stamp indicating when the entry is posted.

There are various links among blogs and entries. A blog page may contain links to archives of old entries. It may also contain a *blogroll*—a sidebar consisting of bookmarks pointing to other blog sites. In the content of an entry, there may be *citation* links pointing to Web sites (e.g., sources of information discussed in the entry) or other entries (written either by the same author or by other bloggers). At the end of an entry, there may be *comments* from other bloggers as well as *trackbacks*—links to other bloggers who are interested in the entry. In this study we mainly focused on the entry-to-entry citation links, i.e., the links in entry contents that point to other entries.

## 2.2 Related Work

There exists a growing body of literature on trend analysis in the blogosphere. In one of the first research papers focusing on blogs, Kumar et al. [13] studied the evolution of the blogosphere. In their study, Kumar et al. studied the evolution of the blog as a graph in terms of the change of characteristics (such as in-degree, out-degree, strongly connected components), the change of communities, as well as the burstiness in blog community. Glance et al. [7] described the BlogPulse system that automatically discovers trends from blogs. In the BlogPulse system, *hot topics*, in the form of key phrases, key persons and key paragraphs are extracted from the blogs. Gruhl et al. [9] studied information diffusion in the blogosphere. and categorized topic propagation into *just spike*, *spiky chatter*, and *mostly chatter*. Our research differs from these studies in two ways. First, these studies based either only on content information in blogs ([7, 9]) or only on link analysis ([13]), while our work combines the contents and the links among blogs. Second, these studies only provided trends in terms of counts (either phrase counts [7, 9] or link counts [13]), while the new techniques we proposed reveal much more information other than the change of counts in the blogosphere.

We have just learned about the recent work by Mei et al. [17]. In this work, time and location of blogs are taken into consideration and spatio-temporal theme patterns are mined by using a probabilistic approach. Our method follows similar ideas (on temporal trends only) but adopts a non-probabilistic approach that is based on singular value decompositions. It will be very interesting to compare the results from the two different approaches.

Modeling the content and linkage changes in graph structures is another research direction that is related to our work. Kumar et al. [13] studied the bursty evolution of links among different communities in the blogosphere by using a Markov chain model. Gruhl et al. [9] proposed a generative model—transmission graph—to model the information diffusion in the blogosphere in a way similar to modeling disease-propagation in epidemic studies. Leskovec et al. [16] studied the patterns of growth for graphs in various fields and proposed generators that produce graphs exhibiting the discovered patterns. Song et al. [20] proposed to represent the expertise of an author by the citation graph of the author’s papers, where an exponential random graph model is

applied to the citation graph to characterize the expertise and the evolution of expertise of the author. Compared to these studies that tried to use various models to fit and explain data, our study focuses instead on direct analysis of the data in order to reveal trends and other insights about the data.

Analyzing the evolution of the World Wide Web is another important research area that is closely related to our work. Murray [18] measured the size and the growth rate of the Internet in the year of 2000. Douglis et al. [5] studied the dynamic characteristics of Web pages from a Web server point of view, with a goal of improving the efficiency of Web caching and delta-encoding. Cho et al. [1] studied the Web page evolution over time from a crawler point of view, with a goal of improving the freshness of crawled Web pages. Later, Fetterly et al. [6] and Ntoulas et al. [19] extended Cho’s work to show further statistics—such as the birth, death and replacement rates for the Web pages and links, the level of change for Web content and the temporal correlations among the changes—on the evolution of the Web. However, all the above work focused on the evolution of the general World Wide Web. It is our belief that the blogosphere has become an important special part of the Web that deserves special study. For example, in their experimental studies, Ntoulas et al. [19] found that 8% of the Web pages are replaced every week and they considered this rate of change very significant. However, in the blogosphere, much higher rate of change is not unexpected—as a matter of fact, in the blogosphere, content changes are so common and so fast that special protocols such as RSS and ATOM are needed for syndication.

## 3. TREND ANALYSIS BASED ON SVD

For a given keyword (e.g., the name of a specific product), trend analysis studies how its popularity in the blogosphere changes over time. For simplicity of discussion, we assume that the blogosphere consists of  $m$  blogs and that the popularity score of a keyword  $k$  among those blogs within a time window  $j$  is given as a popularity vector  $\vec{x}_j = (x_{1j}, \dots, x_{mj})^T$ . This popularity vector is observed through  $n$  consecutive time windows and stacked into an  $m \times n$  matrix  $X = (\vec{x}_1, \dots, \vec{x}_n)$ . Our discussion is independent of how the popularity score is derived. In this paper we take one example:  $x_{ij}$  is the number of entries by blog  $i$  that contain keyword  $k$  at time  $j$ . Then our problem is defined as follows: Given a keyword  $k$ , find a trend vector  $\vec{t} = (t_1, \dots, t_n)^T$  that represents the temporal aspect of the observed popularity scores  $X$ , where  $t_j$  represents the overall popularity score at time  $j$ .

Our idea is representing the observed data  $X$  with a pair of vectors: a trend vector  $\vec{t}$  that represents the overall trends over time and an authority vector  $\vec{a}$  that represents the contribution of individual bloggers to the trend. In the following mathematical formulation, we will show that this pair of vectors can provide better statistic estimation of the observed data  $X$  compared to traditional count-based methods. Based on this discussion, we propose a new temporal trend, called a scalar eigen-trend.

### 3.1 Mathematical Formulation

First, we will see how well traditional count-based methods can represent the observed data  $X$ . A simple count-based method represents the trend as a vector  $\vec{t}_c = (t_1, \dots, t_n)$

where  $t_j = \sum_i x_{ij}$ . That is, the overall popularity score at time  $j$  is defined as the total number of entries among all blogs at time  $j$  that contain the keyword. This count-based score is a good estimator of the central tendency of the popularity among blogs and it is optimal in the following sense—if we assume that at time  $j$ , each  $x_{ij}$  is an independent sample drawn from a random variable with mean  $\frac{1}{m}\mu$ , then  $\hat{\mu} = t_j = \sum_i x_{ij}$  is an unbiased estimator for  $\mu$  that has the minimal sample variance  $\sum_i (x_{ij} - \frac{1}{m}\mu)^2$ . To represent this property in a different way, the vector  $\vec{t}_c$  is the solution to the following equation

$$\vec{t}_c = \arg \min_{\vec{t}_1} \|X - \vec{a}_o \cdot \vec{t}_1^T\|_F \quad (1)$$

where  $\|\cdot\|_F$  is the Frobenius norm and  $\vec{a}_o$  is a column vector whose entries are all  $\frac{1}{m}$ .

Note that, in the above argument, we ignore differences among individual blogs assuming that the popularity score of any blog has the same distribution. That is, the count-based score is a good estimator without knowledge on individual bloggers *a priori*. In reality, however, we can observe that one blogger publishes entries on the keyword more frequently than others, contributing to the popularity of the keyword (i.e., trend) constantly. For example, for the keyword *iPod*, there can be blogs devoted completely to iPod that have tens of entries every day talking about different features of iPod, and there can also be blogs that mention iPod only infrequently.

Let us assume that we know the fraction of contribution to the trend by individual bloggers, that is, we know that  $x_{ij}$  is drawn from a distribution with  $a_i\mu$  as the mean. This information can be given as a unit 2-norm vector  $\vec{a} = (a_1, \dots, a_m)^T$ . Under this assumption, a better trend indicator can be given as  $\mu$  that minimizes the error  $\sum_i (x_{ij} - a_i\mu)^2$  instead of the error  $\sum_i (x_{ij} - \frac{1}{m}\mu)^2$  as used in the count-based method. Then, the trend  $\vec{t}$  is the solution to the following equation:

$$\vec{t} = \arg \min_{\vec{t}_1} \|X - \vec{a} \cdot \vec{t}_1^T\|_F \quad (2)$$

In fact, the following property, whose proof is given in Appendix A, shows that that under an assumption of equal variance, the solution that minimizes  $\sum_i (x_{ij} - a_i\mu)^2$  is the linear unbiased estimator for  $\mu$  with the minimal variance.

**PROPERTY 1.** *Let  $\vec{a} = (a_1, \dots, a_m)^T$  be a unit vector. If for each  $i$ ,  $x_{ij}$  is drawn from a distribution with mean  $a_i\mu$  and variance  $\sigma^2$ , then the value  $\hat{\mu} = \arg \min_r \sum_i (x_{ij} - a_i r)^2$  is the linear unbiased estimator for  $\mu$  with the minimal variance.*

Now, a question is how to estimate  $\vec{a}$ . A simple way is to take the average of  $x_{ij}$  over all the time windows. However, this estimation treats all the time windows equally. Similarly to the above discussion, if we know the trend for each time window, a better way to estimate is to find  $\vec{a}$  that minimizes the error  $\sum_{ij} (x_{ij} - a_i t_j)^2$ . Note that  $\vec{t}$  is the trend we want to find in the first place. Then the trend  $\vec{t}$  is given by the following equation:

$$\vec{t} = \arg \min_{\vec{t}_1} \left( \min_{\|\vec{a}_1\|=1} \|X - \vec{a}_1 \cdot \vec{t}_1^T\|_F \right) \quad (3)$$

That is, we want a pair of  $\vec{t}$  and  $\vec{a}$  that together best approximate the observed data.

Equation (3) can be solved by applying the singular value decomposition on  $X$ :

**THEOREM 1.** *Assume  $X = U\Sigma V^T$  is the singular value decomposition for  $X$ , where  $U = (\vec{u}_1, \dots, \vec{u}_m) \in \mathbb{R}^{m \times m}$  and  $V = (\vec{v}_1, \dots, \vec{v}_n) \in \mathbb{R}^{n \times n}$  are orthogonal matrices representing the basis for the column space and the basis for the row space of  $X$ , respectively;  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0) \in \mathbb{R}^{m \times n}$  in which  $k \leq \min(m, n)$  is the rank of  $X$  and  $\sigma_1 \geq \dots \geq \sigma_k \geq 0$  are the singular values of  $X$ . Then  $\sigma_1 \vec{v}_1$  is a solution to  $\vec{t}$  in Equation (3) and the minimal error is achieved at  $\vec{a}_1 = \vec{u}_1$ .*

**PROOF.** The theorem can be obtained from the following well-known property of the SVD [8]: with  $\sigma_1$ ,  $\vec{u}_1$  and  $\vec{v}_1$  being the first singular value, the first left and right singular vectors, respectively, if we define  $X_1 = \vec{u}_1 \cdot \sigma_1 \vec{v}_1^T$ , then  $\|X - X_1\|_F = \min_{\text{rank}(Y)=1} \|X - Y\|_F$ .

Obviously  $\vec{a}_1 \cdot \vec{t}_1^T$  is a rank-1 matrix with  $\|\vec{a}\| = 1$ . So by taking  $\vec{t}_1 = \sigma_1 \vec{v}_1$  and  $\vec{a}_1 = \vec{u}_1$ , Equation (3) is satisfied.  $\square$

### 3.2 Scalar Eigen-Trend

The above discussion shows that the pair of  $\vec{t}$  and  $\vec{a}$  is a good indicator that approximates the observed data, where the former shows the temporal trend of the popularity and the latter shows the contribution of individual bloggers to the trend. We name them an *eigen-trend* and *authority scores*, respectively. In order to distinguish this trend from another trend we will present later, we specifically call it a *scalar eigen-trend*. These names are given because of the following property:

**PROPERTY 2.** *It can be shown that the solutions  $\vec{a}$  and  $\vec{t}$  from the above procedure satisfies the following recursive relationship (after appropriate normalization)*

$$\begin{cases} \vec{t} &= X^T \vec{a} \\ \vec{a} &= X \vec{t} \end{cases} \quad \text{or} \quad \begin{cases} t_j &= \sum_i x_{ij} a_i \\ a_i &= \sum_j x_{ij} t_j \end{cases} \quad (4)$$

This mutual reinforcement relationship between  $\vec{t}$  and  $\vec{a}$  is similar to the one between *hubs* and *authorities* in the HITS algorithm [11], where a good *hub* page is defined as a page that refers to many good *authority* pages, and a good *authority* page is defined as a page that is referred to by many good *hub* pages. In our case, a blog  $i$  that has a high score  $a_i$  can be seen as an *authority* in a sense that the blogger better represents the trend: the overall popularity  $t_j$  at time  $j$  is high when it is contributed by many good authority blogs, and a good authority blog must contribute to the popularity when the overall popularity  $t_j$  is high.

The scalar eigen-trend and authority scores also have the following properties:

**PROPERTY 3.** *If all elements of  $X$  are non-negative, then the singular value decomposition can be written in such a way that all elements of  $\vec{u}_1$  (and therefore  $\vec{a}$ ) and  $\vec{v}_1$  (and therefore  $\vec{t}$ ) are non-negative.*<sup>6</sup>

**PROPERTY 4.** *When  $\vec{a} \cdot \vec{t}^T$  is used to approximate  $X$ , the square error can be derived from the second through the last singular values as  $\|X - \vec{a} \cdot \vec{t}^T\|_F^2 = \sum_{i>1} \sigma_i^2$ .*

<sup>6</sup>Notice that we can make all elements of  $\vec{u}_1$  and  $\vec{v}_1$  non-positive by flipping the signs of  $\vec{u}_1$  and  $\vec{v}_1$  at the same time.

Property 3 guarantees that  $\vec{a}$  and  $\vec{t}$  to be non-negative. This is helpful because we use  $\vec{t}$  to represent the temporal trend and  $\vec{a}$  to represent the authority score, and it will be difficult to interpret negative values in either of them. Property 4 provides a measure on how much information is captured by the eigen-trend and the authority score.

### 3.3 Benefits of Scalar Eigen-Trend

Compared to traditional count-based trends, the scalar eigen-trend can capture the main stream of blog activity more clearly. In the blogosphere, bloggers publish entries typically driven by events (e.g., press releases of new products). If many bloggers react to the same events at the same time, their synchronous activity forms a “trend.”

The authority score of a blog serves as a “track record” of the blog over time to indicate the amount of contribution to the main-stream trend. An analyst can focus on such authoritative blogs to get deeper insights on the trend. On the other hand, if a blogger behaves independently from the main-stream trend, its authority score is small, and its effect to the trend is discounted. This means that the scalar eigen-trend is less noisy than the count-based trends to extract the main trend from the observed data. We will demonstrate this feature through experiments on synthetic data sets.

The scalar eigen-trend can also capture multiple trends. When the second singular value is large (i.e. the square error of Property 4 is large), another (secondary) trend can be further extracted by using the second singular vector. For example, the same keyword (e.g., tax) can be populated by different groups of blogs that have different points of view (e.g., finance vs. politics). The underlining principle is similar to that of the latent semantic analysis in information retrieval [4]: there can be *latent trends* on the same keyword, which are combined into the observed data. The traditional count-based method is not able to decompose such trends. We will show in the experiments that our method discovers interesting secondary trends from non-dominating interest groups of blogs.

## 4. TREND ANALYSIS BASED ON HOSVD

In the previous section, an element  $x_{ij}$  of matrix  $X$  represents the popularity score of blog  $i$  at time  $j$ . This popularity can simply be measured by the number of relevant entries by blog  $i$  at time  $j$ . However, such a simple definition has a weak point: it ignores the link information in the blogosphere. Blogosphere as a graph structure contains much richer information than a bag of blogs. For example, if relevant entries by a certain blog always attract a lot of links (i.e., references) from other blogs, then this blog should be considered as more important. As another example, because the blogosphere is an ecosystem in which people are mutually aware of each other and interact with each other, we expect that for a given keyword, there exist related communities that exhibit structural consistency over time.

For a given keyword we construct a graph  $G_j$  for time  $j$ , which we call the *keyword-specific blog graph*, as follows. The nodes of  $G_j$  are the  $m$  blogs. There exists an edge  $e_{pq}$  pointing from blog  $b_p$  to blog  $b_q$  if at time  $j$ , there are  $k$  ( $k \geq 1$ ) links pointing from entries in  $b_p$  to entries in  $b_q$  that are related to the keyword. We further set the weight of  $e_{pq}$  to be  $k$ . In this study, we simply define an entry-to-entry link  $e_{pq}$  to be related to a keyword if either the citing entry in  $b_p$  or the cited entry in  $b_q$  contains the keyword. The

keyword-specific blog graph is observed through  $n$  consecutive time windows. If we represent each graph as an  $m \times m$  adjacency matrix, the entire data is represented as a third-order tensor  $\mathcal{X} \in \mathbb{R}^{m \times m \times n}$ , where the first two dimensions of  $\mathcal{X}$  are respectively the rows and columns of the adjacency matrices, and the third dimension is the time line.

In this section we propose a method that *directly* analyzes trends in dynamically changing graph structures, introducing a new temporal trend, which we call *structural eigen-trend*. Similar to the previous section, we apply higher-order singular value decomposition (HOSVD) to the observed data  $\mathcal{X}$ . We will show that  $\mathcal{X}$  can be represented by a triple of vectors: a trend vector  $\vec{t}$  (i.e., a structural eigen-trend), an authority vector  $\vec{a}$ , and a hub vector  $\vec{h}$ . Whereas the previous analysis on scalar eigen-trends represents the characteristics of individual bloggers with a single vector (i.e., an authority vector), this new analysis provides a pair of vectors  $\vec{a}$  and  $\vec{h}$ . As we will discuss, this pair is closely related to the hub and authority in HITS algorithm. In this sense, our analysis captures a *community* that consists of hub and authority blogs and tracks the structure of the community over time.

### 4.1 Mathematical Formulation

We apply singular value decomposition to  $\mathcal{X}$  for trend analysis on the dynamically changing graph structure. However, unlike the case of a matrix, singular value decomposition cannot be uniquely defined on higher-order tensors. Among various techniques developed [2, 12], we adopt the framework proposed by De Lathauwer et al. [2], which is described as follows.

First the singular value decomposition  $X = U\Sigma V^T$  can be rewritten by using *n-mode product* as:

$$X = \Sigma \times_1 U \times_2 V \quad (5)$$

where in general, for a tensor  $\mathcal{A} \in \mathbb{R}^{I_1 \times \dots \times I_n \times \dots \times I_N}$ , the *n-mode product operator*  $\times_n$  of  $\mathcal{A}$  by a matrix  $M \in \mathbb{R}^{J_n \times I_n}$  will result in a tensor  $\mathcal{B} = \mathcal{A} \times_n M \in \mathbb{R}^{I_1 \times \dots \times I_{n-1} \times J_n \times I_{n+1} \times \dots \times I_N}$  where

$$(\mathcal{B})_{i_1 \dots i_{n-1} j_n i_{n+1} \dots i_N} = \sum_{i_n=1}^{I_n} (M)_{j_n i_n} (\mathcal{A})_{i_1 \dots i_{n-1} i_n i_{n+1} \dots i_N}$$

In other words, an *n-mode product*  $\times_n$  of  $\mathcal{A}$  is to apply a linear transformation (represented by  $M$ ) to all the *n-mode vectors* of  $\mathcal{A}$ , where an *n-mode vector* of  $\mathcal{A}$  is an  $I_n$ -dimensional vector obtained by varying the *n*th index of  $\mathcal{A}$  from 1 to  $I_n$  while keeping all other indices fixed.

Because a matrix is a special case of tensor, the natural question is if we can generalize Equation (5) to singular value decomposition on higher-order tensors. De Lathauwer et al. [2] proposed a way of doing that and they called their method the higher-order singular value decomposition (HOSVD). De Lathauwer et al. showed that for a tensor  $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ , we can decompose  $\mathcal{X}$  as

$$\mathcal{X} = \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \dots \times_N U^{(N)} \quad (6)$$

where  $U^{(n)} \in \mathbb{R}^{I_n \times I_n}$  are orthogonal matrices. In Equation (6),  $\mathcal{S} \in \mathbb{R}^{I_1 \times \dots \times I_N}$  is called the *core tensor*. In general,  $\mathcal{S}$  is not diagonal (in the sense that non-zero elements only occur at positions where  $i_1 = \dots = i_N$ ) and the decomposition given by Equation (6) does not have the property of best low-rank approximation. However, De Lathauwer et al.

further proposed an iterative power method that guarantees the best rank-1 approximation [3].

Based on this power method, we propose the following steps to compute the trend in the blogosphere. We first build the third-order tensor  $\mathcal{X}$  as described above to represent the dynamic change of the keyword-specific blog graph over time. Then we use the following iterative method (with appropriate normalization) to compute the first left ( $\vec{h}$ ), right ( $\vec{a}$ ), and third-mode ( $\vec{\tau}$ ) singular vectors.

$$\begin{cases} \vec{h}_{k+1} &= \mathcal{X} \times_2 \vec{a}_k \times_3 \vec{\tau}_k; \\ \vec{a}_{k+1} &= \mathcal{X} \times_1 \vec{h}_{k+1} \times_3 \vec{\tau}_k; \\ \vec{\tau}_{k+1} &= \mathcal{X} \times_1 \vec{h}_{k+1} \times_2 \vec{a}_{k+1}; \\ \lambda_{k+1} &= \|\vec{\tau}_{k+1}\|; \end{cases} \quad (7)$$

It can be shown that the above iteration converges to solutions  $\vec{h}, \vec{a}, \vec{\tau}, \lambda$  such that  $\vec{h} \circ \vec{a} \circ \lambda \vec{\tau}$ , with  $\circ$  being the tensor outer product, is the rank-1 tensor that best approximates  $\mathcal{X}$  in terms of Frobenius norm (square error). In our method, we use  $\vec{t} = \lambda \vec{\tau}$  to represent the temporal trend for the keyword-specific blog graph.

## 4.2 Structural Eigen-Trend

We call the above new trend  $\vec{t}$  a *structural eigen-trend* to distinguish it from the *scalar eigen-trend*. We call the first left and right singular vectors,  $\vec{h}$  and  $\vec{a}$ , *hub scores* and *authority scores*, respectively, based on the following intuitive interpretations.

In the HITS algorithm, for an adjacency matrix  $X$ , the *hub score*, which is the first left singular vector of  $X$ , represents the goodness of the Web pages on summarizing the keyword; the *authority score*, which is the first right singular vector of  $X$ , represents the goodness of the Web pages on being authorities of the keyword. In our case, because  $\vec{h}$  and  $\vec{a}$  are extracted from the tensor  $\mathcal{X}$ , they can be considered as the general hub and authority scores that capture the main community structure related to the keyword in the dynamically changing keyword-specific blog graph. From Equation (7) we can see that after  $\vec{h}$  and  $\vec{a}$  have converged, the trend at time  $j$  is the projection of the keyword-specific blog graph  $G_j$  onto the main community represented by the outer product of  $\vec{h}$  and  $\vec{a}$ . Also from Equation (7) we can see the following:

**PROPERTY 5.** *The HITS algorithm is a special case of our method by taking a single time window, i.e., taking  $n$  to be 1.*

We would also like to point out that although we have used the keyword-specific blog graph as an example, the trend analysis technique presented in this section can be applied to other general graph structures. Furthermore, the same technique can be applied to analyze dynamically changing *undirected* graph structures. In the cases of undirected graphs, instead of the pair of hub and authority scores, we will have a single eigenvector that represents the main “shape” of the graph structures.

In addition, the following property for the trend analysis based on HOSVD can be easily verified.

**PROPERTY 6.** *If all elements of a third-order tensor  $\mathcal{X}$  are non-negative, the iteration given in Equation (7) will converge to a solution such that  $\vec{h}, \vec{a}, \vec{\tau}$ , and  $\lambda$  are all non-negative.*

## 4.3 Benefits of Structural Eigen-Trend

Compared to the scalar eigen-trends, the structural eigen-trends developed in this section focus on and exploit the link structure in the blogosphere. Whereas the scalar eigen-trends emphasize the main group of blogs that publish entries individually, the structural eigen-trends depict activity of the main community that consists of hubs and authorities referencing each other.

In contrast to just applying the HITS algorithm to individual time windows, our new analysis tracks linking behavior of the blogs to find constant hubs and authorities over time. It can discount effects from a blog that does not follow the main trend on linking behavior (for example, a blog that generates links randomly) even if it looks like a hub within a specific time window. Experiments on synthetic data sets will demonstrate that we can focus on the main structural trend and remove noisy behavior. We will also describe interesting findings in the real data set through our analysis.

Similar to the scalar eigen-trend, the secondary trend can be useful to detect another community behaving differently from the main community.

## 5. EXPERIMENTAL RESULTS

In this section, we demonstrate our trend analysis through experiments. First, we conduct experiments on synthetic data sets to verify the benefits of eigen-trends described in Sections 3 and 4. Then, we give case studies on a real blog data set to show interesting trends that are revealed by our methods, which are not available through traditional count-based methods.

### 5.1 Synthetic Data Set

The synthetic data sets are generated as follows. To study the SVD-based trend extraction method, we generate entries from 10 blogs over 250 time units. In a time unit, each blog generates a random number of entries where the number follows a uniform distribution. The mean values of the distribution are different for different blogs. For easy viewing, we let the mean values vary with time following a sinusoid trend.

To study the HOSVD-based trend extraction method, we generate links among 10 blogs over 250 time units. The number of links in each time unit follows a uniform random distribution whose mean value varies over time following a sinusoid trend. When a link is generated, unless stated otherwise, a source blog and a target blog are selected at random, following distributions pre-defined by two unit vectors. These two vectors serve as the underlining hub and authority scores. Compared with the real blogosphere, the scale of our toy examples is very small. However, we use these small examples to illustrate the benefits of our trend analysis methods in certain ideal cases.

#### 5.1.1 Scalar Eigen-Trend

**Case 1:** In this example, the data set is generated in such a way that two blogs (blogs 2 and 8) dominate the entries. That is, when generating entries, the mean values for the random distributions of blogs 2 and 8 are higher than those of other blogs. This data set simulates the case in which a few blogs dominate the discussion on a topic in the blogosphere (e.g., blogs that are completely devoted to reviewing the features of iPod). Then, at time 90, one of the

dominating blogs, blog 8, generates much fewer entries than usual. The experimental results are given in Figure 2. All our figures for trend (e.g., count-based trend, scalar eigen-trend, and structural eigen-trend) have the  $x$ -axis represents the time windows and the  $y$ -axis represents the trend values. For the singular vectors, the  $x$ -axis denotes the blog index. For the singular values, the  $x$ -axis denotes the index for the singular values. As can be seen from the Figures 2(a) and 2(c), both count-based method and the SVD-based method successfully capture the main sinusoid temporal trend. However, the scalar eigen-trend clearly captures the under-representation of the dominating blog at time 90, whereas in the count-based trend, this drop is much less distinct. In addition, the SVD-based method automatically computes the authorities of all the blogs (the first left singular vector shown in Figure 2(d)) and the measure on the approximation error for the main scalar eigen-trend (the singular values shown in Figure 2(b)).

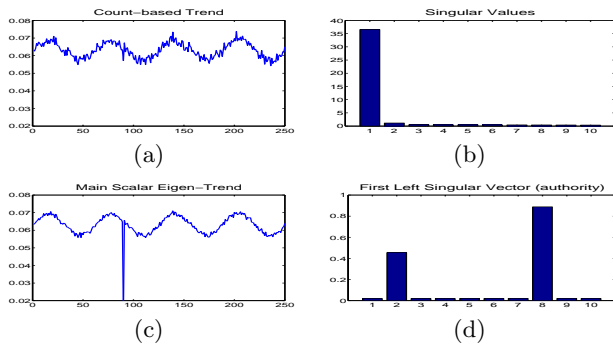


Figure 2: Case 1 – Scalar Eigen-Trend Analysis

**Case 2:** This is an example somewhat contrary to Case 1. In this example, the data set is generated in a similar way to Case 1, but at time 90, one non-dominating blog, blog 5, posts an abnormally large number of entries. As can be seen from the results shown in Figure 3, this abnormality is largely ignored by the scalar eigen-trend in Figure 3(c). In comparison, the count-based trend in Figure 3(a) is impacted greatly. This example illustrates that in scalar eigen-trends, for a blog to have high impact on a keyword, a track record is needed to be built over time, and one-time shot does not count very much.

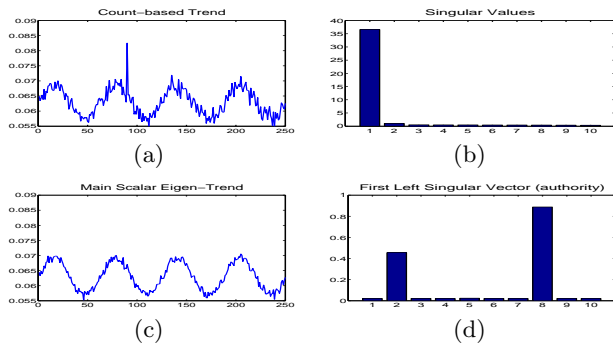


Figure 3: Case 2 – Scalar Eigen-Trend Analysis

**Case 3:** In this example, we illustrate the point of multiple trends. When generating the data set, during the first 150 time units, blogs 2 and 8 dominate the entries and then during the last 100 time units, the dominating blogs are switched to blogs 4 and 6. This example is used to simulate the case in which two distinct groups of blogs discuss different aspects of the same keyword following different temporal patterns. In Figure 4, we show the experimental results for these 2 blog groups. The first and second scalar eigen-trends in Figures 4(c) and 4(e) accurately capture trends in the two interest groups. In addition, the corresponding authority scores (left singular vectors) shown in Figures 4(d) and 4(f) reflect the membership of the blogs in each interest group. Furthermore, the magnitude of the singular values (shown in Figure 4(b)) provides hint on how dominating each group of blogs are in the blogosphere.

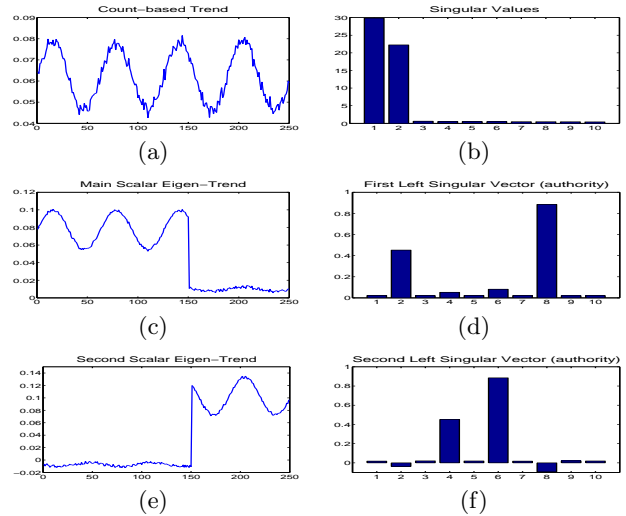


Figure 4: Case 3 – Scalar Eigen-Trend Analysis

### 5.1.2 Structural Eigen-Trend

**Case 4:** In this example, when a link is generated, the probability for a blog to be chosen as the source blog is uniformly distributed among blogs 1,3,5,7, and 9; the probability for a blog to be chosen as the target blog is uniformly distributed over blogs 2 and 8. In addition, random links are added as noise. However, at time 90, the graph structure changes. At time 90, instead of using the hub and authority scores, all links are generated totally randomly by equally likely selecting any blog to be the source or the target. The results are shown in Figure 5. The structural change is detected by the structural eigen-trend in Figure 5(c) but is not detectable by the count-based trend in Figure 5(a). The drop in the structural eigen-trend suggests that at time 90 the number of links that follow the normal graph structure (which is represented as the authority and hub scores) is much lower than usual, which suggests a structural change at time 90.

**Case 5:** This example is somewhat contrary to the above example. Links are generated in a similar way to Case 4. At time 90, blog 6, which is not a good hub, generates a lot of links pointing to the two authorities. Results shown in Figure 6 demonstrate that while this spam-like behavior impacts the count-based trend in Figure 6(a) greatly, the

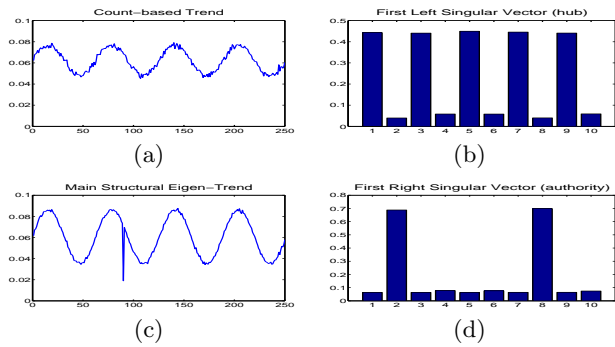


Figure 5: Case 4 – Structural Eigen-Trend Analysis

structural eigen-trend in Figure 6(c) largely ignores these usual links. That is, to become a valid hub, a blog must build a track record of consistently pointing to good authorities over all the time.

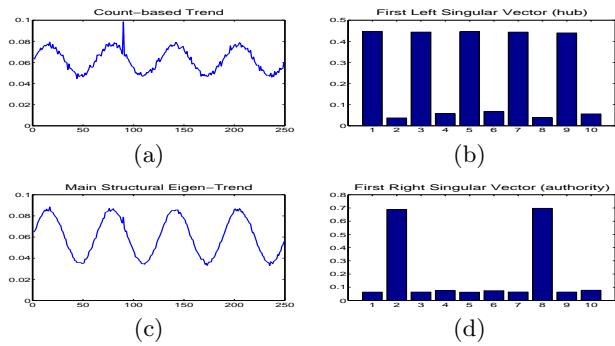


Figure 6: Case 5 – Structural Eigen-Trend Analysis

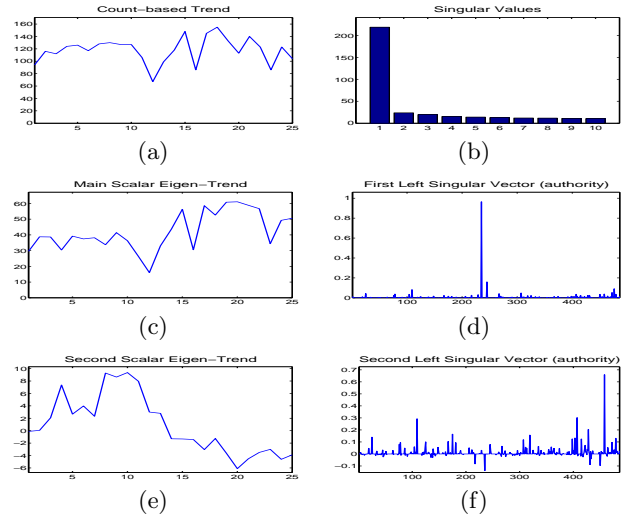
## 5.2 Real Blog Data Set

For real data, we use a blog data set obtained by an in-house crawler developed at NEC Laboratories American. For our analysis, we extracted a subset of English blogs consisting of 114,645 entries that belong to 486 blogs crawled between July 10th and December 30th in 2005, for a period of 25 consecutive weeks. In addition, there are totally 34,994 links in the data set. Although the data set is relatively small compared to those from large-scale commercial blog search engines, we show that our new techniques are able to discover very interesting trends that are not available through traditional methods.

### 5.2.1 Scalar Eigen-Trend

We demonstrate our scalar eigen-trend analysis and the URL's of top authority blogs for the keyword *tax* as shown in Figure 7. We observe that the first and the second scalar eigen-trends in Figures 7(c) and 7(e) follow very different patterns. It turns out that the first scalar eigen-trend is mainly driven by a group of blogs with financial interests. For example, the blog in this group with the top authority belongs to a law professor, Prof. Paul L. Caron, who is a leading tax scholar. Main topics covered by this group of blogs include IRS rules, tax guide for organizations and individuals, etc. As can be expected, the number of entries from

these blogs increases dramatically toward the end of fiscal year, when tax becomes a more important issue. Because most entries from these blogs contain the keyword *tax*, these blogs dominate the blogosphere and the count-based trend follows this main scalar eigen-trend. On the other hand, the authorities in the second interest group are mainly political blogs. Tax-related topics in these blogs include taxation, tax rates, tax cuts and their political consequences. The second scalar eigen-trend in Figures 7(e) reveals another trend that belong to a group behaving differently from the first group.

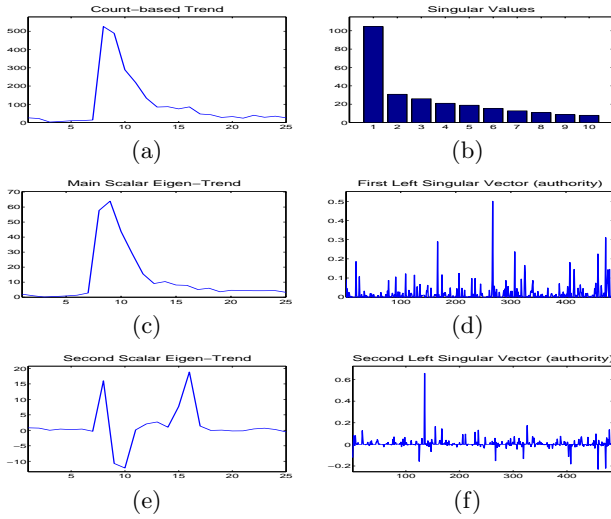


Top Authorities for the First Scalar Eigen-Trend
<a href="http://taxprof.typepad.com/taxprof.blog/">http://taxprof.typepad.com/taxprof.blog/</a>
<a href="http://timworstall.typepad.com/timworstall/">http://timworstall.typepad.com/timworstall/</a>
Top Authorities for the Second Scalar Eigen-Trend
<a href="http://www.theleftcoaster.com/">http://www.theleftcoaster.com/</a>
<a href="http://www.preemptivekarma.com/">http://www.preemptivekarma.com/</a>
<a href="http://ezraklein.typepad.com/blog/">http://ezraklein.typepad.com/blog/</a>

Figure 7: Eigen-Trend Analysis and Top Authorities for Keyword *Tax*

Our second example is on the keyword *hurricane*. Hurricane Katrina took place during the 7th week in our time frame in Figure 8. As can be seen from the count-based trend in Figure 8(a), many entries were posted immediately after Hurricane Katrina and interests on this topic died out gradually after a few weeks. Also can be seen from the figure, the first scalar eigen-trend in Figure 8(c) obtained by our SVD-based method follows this count-based trend closely and they are driven mainly by some well-known news and political blogs. These blogs mainly reported news related to Hurricane Katrina and discussed the economic and political impacts of the hurricane. In comparison, the second interest group mainly consists of less well-known personal blogs. Their main topics related to Hurricane Katrina include personal experiences, helping the victims, making donations, etc. In the second scalar eigen-trend shown in Figure 8(e) that corresponds to this second group of blogs, another spike occurs in the 16th week. The reason for this spike is that due to the nature of this group, they discussed in a similar fashion on a subsequent hurricane, Hurricane Wilma. Because Hurricane Wilma has less dramatic political or eco-

nomic impact than Hurricane Katrina, as we can see from the figure, it is hardly noticeable in the count-based trend.



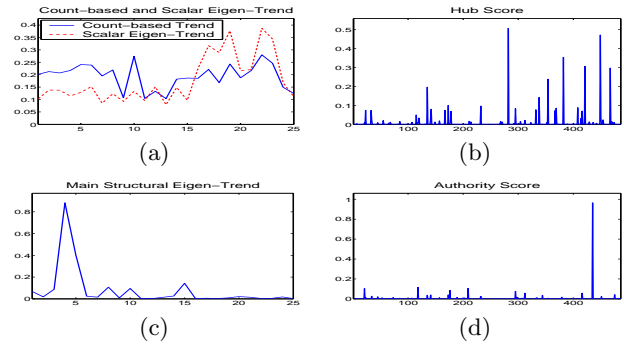
Top Authorities for the First Eigen-Trend
<a href="http://wizbangblog.com/">http://wizbangblog.com/</a>
<a href="http://www.washingtonmonthly.com/">http://www.washingtonmonthly.com/</a>
<a href="http://michellemalkin.com/">http://michellemalkin.com/</a>
Top Authorities for the Second Eigen-Trend
<a href="http://hyku.com/blog/">http://hyku.com/blog/</a>
<a href="http://www.donaldsensing.com/">http://www.donaldsensing.com/</a>
<a href="http://majikthise.typepad.com/majikthise_/">http://majikthise.typepad.com/majikthise_/</a>

Figure 8: Eigen-Trend Analysis and Top Authorities for Keyword *Hurricane*

### 5.2.2 Structural Eigen-Trend

In our experiments, structural eigen-trends extracted by using HOSVD normally comply with trends obtained by using other methods. Here we show an interesting example in which the structural eigen-trend is dramatically different from other trends. Figure 9 shows the result for the keyword *Technorati*, where Technorati is the name of a top blog search company. In the structural eigen-trend of Figure 9(c), there is a large spike at time 4 that is not in either the count-based trend or the scalar eigen-trend of Figure 9(a). Puzzled by this result, we manually checked all the entries that contain the keyword *Technorati* in our data set. It turns out that many of these entries contain a line such as “Technorati Tag: news, music” at the bottom, to indicate the category of the entries. Our crawler failed to remove this line from the body of the entries. As a result, the top authorities for the scalar eigen-trend are some blogs who posted a lot of entries that adopted Technorati tags.

Let us look at the structural eigen-trend. The dominating authority for the structural eigen-trend turns out to be the personal blog site of David Sifry, the founder and CEO of Technorati Inc. In the first week of August 2005 (which is the 4th week in our data set), David Sifry posted the first three parts of a study on the current state of the blogosphere. In this study, based on the data collected by the Technorati search engine, David Sifry presented a lot of statistics and insights about the blogosphere, including the growth of blog, the change of posting volume, and the trend of people adopt-



Top Authorities for Scalar Eigen-Trend
<a href="http://www.ratcliffeblog.com/">http://www.ratcliffeblog.com/</a>
<a href="http://www.emergencemarketing.com/">http://www.emergencemarketing.com/</a>
<a href="http://www.tomrafteryit.net/">http://www.tomrafteryit.net/</a>
Top Hubs for Structural Eigen-Trend
<a href="http://www.ballpark.ch/blog/">http://www.ballpark.ch/blog/</a>
<a href="http://www.techcrunch.com/">http://www.techcrunch.com/</a>
<a href="http://www.morganmclintic.com/pr/">http://www.morganmclintic.com/pr/</a>
Top Authority for Structural Eigen-Trend
<a href="http://www.sifry.com/alerts/">http://www.sifry.com/alerts/</a>

Figure 9: Eigen-Trend Analysis and Top Authorities for Keyword *Technorati*

ing tags in their blogs. Because this was one of the most authoritative studies on the current state of the blogosphere, this study drew a lot of attentions and generated intensive citations. This event is actually visually detectable from Figure 10(a), in which we visualize the adjacency matrices for the keyword-specific blog graph (on *Technorati*) in the first 10 weeks (the arrow in the 4th week points to the blog by David Sifry). However, because of the large number of entries that contain *Technorati* (e.g., by using the Technorati Tag line), neither count-based trend nor scalar eigen-trend is able to detect this important event. In the method based on HOSVD, those blogs that incidently contain *Technorati* do not form a well-structured community and therefore are treated more as noise. In contrast, the community formed by David Sifry’s blog, as well as its followers, form a consistent community (David Sifry has continued posting a sequence of highly cited entries about Technorati in the following weeks). In our HOSVD-based method, this community (which is visualized in Figure 10(b)) stands out as the main community on *Technorati* and as shown in Figure 9, events within this community determine the main structural eigen-trend.

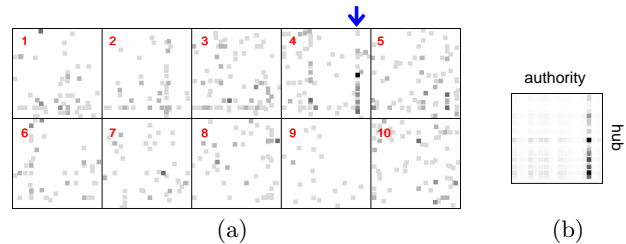


Figure 10: (a) Adjacency Matrices in the First 10 Time Windows, (b) Main Community for *Technorati*

## 6. CONCLUSION AND FUTURE WORK

In this study, we introduced a novel concept—eigen-trend, and based on this new concept, we proposed two novel techniques for trend extraction and analysis in the blogosphere. The first technique, which is based on the singular value decomposition, has good statistical features and can reveal multiple trends from different groups of blogs with different interests. The second technique, which is based on a higher-order singular value decomposition, analyzes the blogosphere as a dynamical graph structure and is able to detect structural changes in the blogosphere. Experimental results on synthetic data sets showed very interesting features of our new methods and case studies on real blog data set demonstrated that our new techniques are able to discover interesting trends that are not available in traditional count-based methods.

For future work, we plan to extend our study into several directions. First, in this study we focused on analyzing historic data; we plan to extend our techniques to predicting future trends. Second, the computations in this study were all off-line; we plan to develop on-line incremental algorithms to make our techniques applicable to real-time systems. Third, the trend analysis in this paper was based on individual keywords; in the future, we plan to study how to combine trends for multiple keywords so that our techniques can be applied to other temporal problems, such as community evolution and information diffusion, in the blogosphere. Finally, it is computationally expensive to compute SVD and HOSVD. Although our current implementation is practical for small blog data sets (e.g., results of sampling or querying the blogosphere), for larger data we do expect scalability to be a challenging issue. To solve this issue, we plan to improve our implementation by using incremental algorithms or by taking advantage of the sparseness of the blogosphere.

## 7. REFERENCES

- [1] J. Cho and H. Garcia-Molina. Effective page refresh policies for web crawlers. *ACM Tran. on Database Systems*, 28(4), 2003.
- [2] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. on Matrix Analysis and Applications*, 21(4), 2000.
- [3] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank- $(r_1, r_2, \dots, r_n)$  approximation of higher-order tensors. *SIAM J. on Matrix Analysis and Applications*, 21(4), 2000.
- [4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *J. American Soc. Info. Sci.*, 41, 1990.
- [5] F. Douglass, A. Feldmann, and B. Krishnamurthy. Rate of change and other metrics: a live study of the World Wide Web. In *Proc. of the USENIX Symposium on Internet Technologies and Systems*, 1997.
- [6] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener. A large-scale study of the evolution of web pages. In *Proc. of the 12th WWW Conference*, 2003.
- [7] N. S. Glance, M. Hurst, and T. Tomokiyo. BlogPulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.
- [8] G. Golub and C. V. Loan. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996.

- [9] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. of the 13th WWW Conference*, 2004.
- [10] I. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002.
- [11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. of the ACM*, 46(5), 1999.
- [12] T. G. Kolda, B. W. Bader, and J. P. Kenny. Higher-order web link analysis using multilinear algebra. In *Proc. of the 5th ICDM Conf.*, 2005.
- [13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of the 12th WWW Conference*, 2003.
- [14] D. Lai. Temporal analysis of the human development indicators: Principal component approach. *Social Indicators Research*, 51, 2000.
- [15] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. In *Proc. of the 2004 SIGMETRICS Conf.*, 2004.
- [16] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of the 11th ACM SIGKDD Conference*, 2005.
- [17] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proc. of the 15th WWW Conference*, 2006.
- [18] B. H. Murray. Sizing the internet. In *White paper, Cyveillance, Inc.*, 2000.
- [19] A. Ntoulas, J. Cho, , and C. Olston. What's new on the Web? the evolution of the web from a search engine perspective. In *Proc. of the 13th WWW Conference*, 2004.
- [20] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun. ExpertiseNet: Relational and evolutionary expert modeling. In *Int. Conf. on User Modeling*, 2005.

## APPENDIX

### A. PROOF OF PROPERTY 1

PROOF. By setting the derivative of  $\sum_i (x_{ij} - a_i r)^2$  with respect to  $r$  to be zero, we have that the value that minimizes  $\sum_i (x_{ij} - a_i r)^2$  is  $\hat{\mu} = \sum_i a_i x_{ij}$ .  $\hat{\mu}$  is an unbiased estimation of  $\mu$  because  $E(\hat{\mu}) = E(\sum_i a_i x_{ij}) = \sum_i a_i^2 \mu = \mu$ . Now we prove that  $\hat{\mu}$  is the linear unbiased estimator for  $\mu$  with the minimal variance.

For an arbitrary linear estimator  $\hat{\mu}_1$  for  $\mu$ , we write  $\hat{\mu}_1$  as  $\sum_i b_i x_{ij}$  and define  $\vec{b} = (b_1, \dots, b_m)^T$ . For  $\hat{\mu}_1$  to be unbiased, we have  $\mu = E(\hat{\mu}_1) = E(\sum_i b_i x_{ij}) = (\sum_i b_i a_i) \mu$  and so  $\sum_i b_i a_i = 1$  or equivalently

$$\|\vec{b}\| \cdot \|\vec{a}\| \cdot \cos\theta = 1 \quad (8)$$

where  $\theta$  is the angle between  $\vec{b}$  and  $\vec{a}$ .

The variance of  $\hat{\mu}_1$  can be written as

$$\text{var}(\hat{\mu}_1) = \text{var}\left(\sum_i b_i x_{ij}\right) = \left(\sum_i b_i^2\right) \sigma^2 = \|\vec{b}\|^2 \sigma^2.$$

So we want to minimize  $\|\vec{b}\|^2 \sigma^2$  subjected to (8). Because  $\|\vec{a}\| = 1$ , the solution is obviously  $\theta = 0$  and  $\vec{b} = \vec{a}$ . Therefore,  $\hat{\mu} = \sum_i a_i x_{ij}$  is the linear unbiased estimator for  $\mu$  with the minimal variance.  $\square$