Chapter 12
Blog Data Mining for Cyber Security Threats

Flora S. Tsai and Kap Luk Chan

Abstract Blog data mining is a growing research area that addresses the domain-specific problem of extracting information from blog data. In our work, we analyzed blogs for various categories of cyber threats related to the detection of security threats and cyber crime. We have extended the Author-Topic model based on Latent Dirichlet Allocation for identify patterns of similarities in keywords and dates distributed across blog documents. From this model, we visualized the content and date similarities using the Isomap dimensionality reduction technique. Our findings support the theory that our probabilistic blog model can present the blogosphere in terms of topics with measurable keywords, hence aiding the investigative processes to understand and respond to critical cyber security events and threats.

12.1 Introduction

Organizations and governments are becoming vulnerable to a wide variety of security breaches against their information infrastructure. The severity of this threat is evident from the increasing rate of cyber attacks against computers and critical infrastructure. According to Sophos latest report, one new infected webpage is discovered every 14 seconds, or 6,000 a day [17].

As the number of cyber attacks by persons and malicious software are increasing rapidly, the number of incidents reported in blogs are also on the rise. Blogs are websites where entries are made in a reverse chronological order. Blogs may provide up-to-date information on the prevalence and distribution of various security incidents and threats.

Blogs range in scope from individual diaries to arms of political campaigns, media programs, and corporations. Blogs' explosive growth is generating large vol-

Flora S. Tsai, Kap Luk Chan
Nanyang Technological University, Singapore, e-mail: fst1@columbia.edu, eklchan@ntu.edu.sg

umes of raw data and is considered by many industry watchers one of the top ten trends. Blogosphere is the collective term encompassing all blogs as a community or social network. Because of the huge volume of existing blog posts and their free format nature, the information in the blogosphere is rather random and chaotic, but immensely valuable in the right context. Blogs can thus potentially contain usable and measurable information related to security threats, such as malware, viruses, cyber blackmail, and other cyber crime.

With the amazing growth of blogs on the web, the blogosphere affects much in the media. Studies on the blogosphere include measuring the influence of the blogosphere [6], analyzing the blog threads for discovering the important bloggers [13], determining the spatiotemporal theme pattern on blogs [12], focusing the topic-centric view of the blogosphere [1], detecting the blogs growing trends [7], tracking the propagation of discussion topics in the blogosphere [8],searching and detecting topics in business blogs [3], and determining latent friends of bloggers [16].

Existing studies have also focused on analyzing forums, news articles, and police databases for cyber threats [14, 21–23], but few have looked at blogs. In this paper, we focus on analyzing security blogs, which are blogs providing commentary or analysis of security threats and incidents.

In this paper, we propose blog data mining techniques for evaluating security threats related to the detection of cyber attacks, cyber crime, and information security. Existing studies on intelligence analysis have focused on analyzing news or forums for security incidents, but few have looked at blogs. We use probabilistic methods based on Latent Dirichlet Allocation to detect keywords from security blogs with respect to certain topics. We then demonstrate how this method can present the blogosphere in terms of topics with measurable keywords, hence tracking popular conversations and topics in the blogosphere. By applying a probabilistic approach, we can improve information retrieval in blog search and keywords detection, and provide an analytical foundation for the future of security intelligence analysis of blogs.

The paper is organized as follows. Section 2 reviews the related work on intelligence analysis and extraction of useful information from blogs. Section 3 defines the attributes of blog documents, and describes the probabilistic techniques based on Latent Dirichlet Allocation [2], Author-Topic model [18], and Isomap algorithm [19] for mining and visualization of blog-related topics. Section 4 presents experimental results, and Section 5 concludes the paper.

12.2 Review of Related Work

This section reviews related work in developing security intelligence analysis and extraction of useful information from blogs.

12.2.1 Intelligence Analysis

Intelligence analysis is the process of producing formal descriptions of situations and entities of strategic importance [20]. Although its practice is found in its purest form inside intelligence agencies, such as the CIA in the United States or MI6 in the UK, its methods are also applicable in fields such as business intelligence or competitive intelligence.

Recent work related to security intelligence analysis include using entity recognizers to extract names of people, organizations, and locations from news articles, and applying probabilistic topic models to learn the latent structure behind the named entities and other words [14]. Another study analyzed the evolution of terror attack incidents from online news articles using techniques related to temporal and event relationship mining [22]. In addition, Support Vector Machines were used for improving document classification for the insider threat problem within the intelligence community by analyzing a collection of documents from the Center for Nonproliferation Studies (CNS) related to weapons of mass destruction [23]. Another study analyzed the criminal incident reporting mainframe system (RAMS) data set used by the police department in Richmond, VA to analyze and predict the spatial behavior of criminals and latent decision makers [21]. These studies illustrate the growing need for security intelligence analysis, and the usage of machine learning and information retrieval techniques to provide such analysis. However, much work has yet to be done in obtaining intelligence information from the vast collection of blogs that exist throughout the world.

12.2.2 Information Extraction from Blogs

Current blog text analysis focuses on extracting useful information from blog entry collections, and determining certain trends in the blogosphere. NLP (Natural Language Processing) algorithms have been used to determine the most important keywords and proper names within a certain time period from thousands of active blogs, which can automatically discover trends across blogs, as well as detect key persons, phrases and paragraphs [7]. A study on the propagation of discussion topics through the social network in the blogosphere developed algorithms to detect the long-term and short-term topics and keywords, which were then validated with real blog entry collections [8]. On evaluating the suitable methods of ranking term significance in an evolving RSS feed corpus, three statistical feature selection methods were implemented: χ^2, Mutual Information (MI) and Information Gain (I), and the conclusion was that χ^2 method seems to be the best among all, but full human classification exercise would be required to further evaluate such method [15]. A probabilistic approach based on PLSA was proposed in [12] to extract common themes from blogs, and also generate the theme life cycle for each given location and the theme snapshots for each given time period. PLSA has also been previously used for blog search and mining of business blogs [3]. Latent Dirichlet Allocation

(LDA) [2] was used for identifying latent friends, or people who share similar topic distribution in their blogs [16].

12.3 Probabilistic Techniques for Blog Data Mining

This section summarizes the attributes of blog documents that distinguish them from other types of documents such as Web documents. The multiple dimensions of blogs provide a rich medium from which to perform blog data mining. The technique of Latent Dirichlet Allocation (LDA) extended for blog data mining is described. We propose a Date-Topic model based on the Author-Topic model for LDA that was used to analyze the blog dates in our dataset. Visualization is performed with the aid of the Isomap dimensionality reduction technique, which allows the content and date similarities to be easily visualized.

12.3.1 Attributes of Blog Documents

A blog document is structured differently from a typical Web document. Table 12.1 provides a comparison of facets of blog and Web documents. URL stands for the Uniform Resource Locator, that describes the Web address from which a document can be found. A permalink is specific to blogs, and is a URL that points to a specific blog entry after the entry has passed from the front page into the blog archives. Outlinks are documents that are linked from the blog or Web document. Tags are labels that people use to make it easier to find blog posts, photos and videos that are related. One important distinction in blog documents that makes them very different from Web documents are the time and date components.

Table 12.1 Comparison of blog and Web documents

Components	**Blog**	**Web**
title	√	√
content	√	√
tags	√	
author	√	
URL	√	√
permalink	√	
outlinks	√	√
time	√	
date	√	

If we consider the different facets of blogs, we can group general blog data analysis into five main attributes (blog content, tags, authors, time, and links), shown in Table 12.2. Each of the attributes itself can be multidimensional.

Table 12.2 Blog attributes

Attributes	**Blog Components**
Content	*title and content*
Tags	*tags*
Author	*author or poster*
Links	*URL, permalink, outlinks*
Time	*date and time*

Another attribute that is not directly present in blogs, but can be extracted from the content or author information, is the blog location, or the geographic location of the blog author. In addition, many blog posts have optional comments for users to add feedback to the blog. Although not part of the original post, comments can provide additional insight into the opinions related to the blog post.

Due to the complexity of analyzing the multidimensional characteristics of blogs, many previous analysis techniques analyze only one or two attributes of the blog data.

12.3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2] is a probabilistic technique which models text documents as mixtures of latent topics, where topics correspond to key concepts presented in the corpus. LDA is not as prone to overfitting, and is preferred to traditional methods based on Latent Semantic Analysis (LSA) [5]. For example, in Probabilistic Latent Semantic Analysis (PLSA) [11], the number of parameters grows with the number of training documents, which makes the model susceptible to overfitting.

In LDA, the topic mixture is drawn from a conjugate Dirichlet prior that is the same for all documents, as opposed to PLSA, where the topic mixture is conditioned on each document. In LDA, the steps adapted for blog documents are summarized below:

1. Choose a multinomial distribution ϕ_z for each topic z from a Dirichlet distribution with parameter β.
2. For each blog document b, choose a multinomial distribution θ_b from a Dirichlet distribution with parameter α.
3. For each word token w in blog b, choose a topic t from θ_b.
4. Choose a word w from ϕ_t.

The probability of generating a corpus is thus equivalent to:

$$\iint \prod_{t=1}^{K} P(\phi_t|\beta) \prod_{b=1}^{N} P(\theta_b|\alpha)(\prod_{i=1}^{N_b} \sum_{t_i=1}^{K} P(t_i|\theta)P(w_i|t,\phi))d\theta d\phi \qquad (12.1)$$

An extension of LDA to probabilistic Author-Topic (AT) modeling [18] is proposed for the blog author and topic visualization. The AT model is based on Gibbs sampling, a Markov chain Monte Carlo technique, where each author is represented by a probability distribution over topics, and each topic is represented as a probability distribution over terms for that topic [18].

We have extended the AT model for visualization of blog dates. For the Date-Topic (DT) model, each date is represented by a probability distribution over topics, and each topic represented by a probability distribution over terms for that topic.

For the DT model, the probability of generating a blog is given by:

$$\prod_{i=1}^{N_b} \frac{1}{D_b} \sum_{d} \sum_{t=1}^{K} \phi_{w_i t} \theta_{td} \tag{12.2}$$

where blog b has D_b dates. The probability is then integrated over ϕ and θ and their Dirichlet distributions and sampled using Markov Chain Monte Carlo methods.

The similarity matrices for dates can then be calculated using the symmetrized Kullback Leibler (KL) distance [10] between topic distributions, which is able to measure the difference between two probability distributions. The symmetric KL distance of two probability distributions P and Q is calculated as:

$$\frac{KL(P,Q) + KL(Q,P)}{2} \tag{12.3}$$

where KL is the KL distance given by:

$$KL(P,Q) = \sum (P \log(P/Q)); \tag{12.4}$$

The similarity matrices can be visualized using the Isomap dimensionality reduction technique described in the following section.

12.3.3 Isometric Feature Mapping (Isomap)

Isomap [19] is a nonlinear dimensionality reduction technique that uses Multidimensional Scaling (MDS) [4] techniques with geodesic interpoint distances instead of Euclidean distances. Geodesic distances represent the shortest paths along the curved surface of the manifold . Unlike the linear techniques, Isomap can discover the nonlinear degrees of freedom that underlie complex natural observations [19].

Isomap deals with finite data sets of points in $\mathbb{R}^n$ which are assumed to lie on a smooth submanifold M_d of low dimension $d < n$. The algorithm attempts to recover M given only the data points. Isomap estimates the unknown geodesic distance in M between data points in terms of the graph distance with respect to some graph G constructed on the data points.

Isomap algorithm consists of three basic steps:

1. Find the nearest neighbors on the manifold M, based on the distances between pairs of points in the input space.
2. Approximate the geodesic distances between all pairs of points on the manifold M by computing their shortest path distances in the graph G.
3. Apply MDS to matrix of graph distances, constructing an embedding of the data in a d-dimensional Euclidean space Y that best preserves the manifold's estimated intrinsic geometry [19].

If two points appear on a nonlinear manifold, their Euclidean distance in the high-dimensional input space may not accurately reflect their intrinsic similarity. The geodesic distance along the low-dimensional manifold is thus a better representation for these points. The neighborhood graph G constructed in the first step of allows an estimation of the true geodesic path to be computed efficiently in step two, as the shortest path in G. The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph. The embedding now represents simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths [19].

Isomap is a very useful noniterative, polynomial-time algorithm for nonlinear dimensionality reduction if the data is severely nonlinear. Isomap is able to compute a globally optimal solution, and for a certain class of data manifolds (Swiss roll), is guaranteed to converge asymptotically to the true structure [19]. However, Isomap may not easily handle more complex domains such as non-trivial curvature or topology.

12.4 Experiments and Results

We used probabilistic models for blog data mining on our dataset. Dimensionality reduction was performed with Isomap to show the similarity plot of blog content and dates. We extract the most relevant categories and show the topics extracted for each category. Experiments show that the probabilistic model can reveal interesting patterns in the underlying topics for our dataset of security-related blogs.

12.4.1 Data Corpus

For our experiments, we extracted a subset of the Nielson BuzzMetrics blog data corpus[1] that focuses on blogs related to security threats and incidents related to cyber crime and computer viruses. The original dataset consists of 14 million blog posts collected by Nielsen BuzzMetrics for May 2006. Although the blog entries span only a short period of time, they are indicative of the amount and variety of blog posts that exists in different languages throughout the world.

[1] `http://www.icwsm.org/data.html`

Blog entries in the English language related to security threats such as malware, cyber crime, computer virus, encryption, and information security were extracted and stored for use in our analysis. Figure 12.1 shows an excerpt of a blog post related to a security glitch found on voting machines.

> Elections officials ... are scrambling to understand and limit the risk from a “dangerous" security hole found in ... touch-screen voting machines. ... Armed with a little basic knowledge of Diebold voting systems ... someone ... could load virtually any software into the machine and disable it, redistribute votes or alter its performance...

Fig. 12.1 Excerpt of blog post related to security glitch in voting machines.

The prevalence of articles and blogs on this matter has led to many proposed legislation reforms regarding electronic voting machines [9]. Thus, security incidents reported in the blogosphere and other online media can greatly effect traditional media and legislation.

There are a total of 2102 entries in our dataset, and each blog entry is saved as a text file for further text preprocessing. For the preprocessing of the blog data, HTML tags were removed, lexical analysis was performed by removing stopwords, stemming, and pruning by the Text to Matrix Generator (TMG) [24] prior to generating the term-document matrix. The total number of terms after pruning and stopword removal is 6169. The term-document matrix was then input to the LDA algorithm.

12.4.2 Results for Blog Topic Analysis

We conducted some experiments using LDA for the blog entries. The parameters used in our experiments are number of topics (10) and number of iterations (1000). We used symmetric Dirichlet priors in the LDA estimation with α = 50/ K and β =0.01, which are common settings in the literature.

Tables 12.3-12.8 summarizes the keywords found for each of the top six topics.

By looking at the various topics listed, we are able to see that the probabilistic approach is able to list important keywords of each topic in a quantitative fashion. The keywords listed can relate back to the original topics. For example, the keywords detected in the Topic 2 include “malwar", “worm", “threat", and “terror". All of these types are related to the general category of computer malware.

For Topic 5, the keywords such as “vote", “machin", “elect", and “diebold" relate to blog posts about security glitches found on voting machines, as shown in the blog summary from Figure 12.1. The high probability of dates around May 12-13 indicate that many of the blog posts occurred during this period of time. These are examples of events that can trigger conversation in the blogosphere.

Automatic topic detection of security blogs such as those demonstrated above can have significant impact on the investigation and detection of cyber threats in the

Table 12.3 List of terms and dates for Topic 1.

Term	**Probability**
comput	0.02956
file	0.01451
click	0.01082
search	0.01063
inform	0.00922
page	0.00922
phone	0.00901
track	0.00846
data	0.00813
record	0.00777
Date	**Probability**
20060513	0.10077
20060512	0.09032
20060503	0.07942
20060505	0.07665
20060516	0.07130
20060502	0.05263
20060507	0.04979
20060514	0.04776
20060523	0.04776
20060504	0.04275

Table 12.4 List of terms and dates for Topic 2.

Term	**Probability**
browser	0.01660
user	0.01444
secur	0.01277
cyber	0.01194
worm	0.01189
comput	0.01130
instal	0.01097
terror	0.01084
malwar	0.01079
threat	0.01073
Date	**Probability**
20060523	0.39632
20060522	0.17909
20060524	0.08151
20060521	0.04653
20060519	0.04547
20060512	0.03673
20060505	0.03573
20060513	0.02941
20060504	0.02635
20060515	0.01843

Table 12.5 List of terms and dates for Topic 3

Term	**Probability**
window	0.01766
encrypt	0.01297
work	0.01182
secur	0.01120
kei	0.01115
network	0.01109
run	0.00955
system	0.00892
server	0.00869
support	0.00751
Date	**Probability**
20060502	0.10822
20060504	0.08676
20060507	0.08144
20060512	0.07670
20060518	0.07605
20060503	0.07308
20060505	0.07183
20060514	0.06587
20060524	0.06260
20060519	0.04735

Table 12.6 List of terms and dates for Topic 4.

Term	**Probability**
privaci	0.01601
servic	0.00971
spywar	0.00924
data	0.00896
inform	0.00849
law	0.00847
part	0.00837
time	0.00792
right	0.00774
power	0.00759
Date	**Probability**
20060519	0.10077
20060521	0.09032
20060513	0.07942
20060505	0.07665
20060512	0.07130
20060518	0.05263
20060522	0.04979
20060524	0.04776
20060523	0.04776
20060504	0.04275

blogosphere. A high incidence of occurrence of a particular topic or keyword can alert the user of potential new threats and security risks, which can then be further

Table 12.7 List of terms and dates for Topic 5.

Term	Probability
vote	0.02172
machin	0.01619
elect	0.01082
state	0.01257
call	0.01155
diebold	0.01106
bush	0.00927
system	0.00912
secur	0.00873
nsa	0.00870
Date	**Probability**
20060512	0.44553
20060513	0.20167
20060515	0.07230
20060511	0.05536
20060514	0.03944
20060505	0.03734
20060523	0.02488
20060519	0.02310
20060518	0.01288
20060504	0.01280

Table 12.8 List of terms and dates for Topic 6.

Term	Probability
secur	0.02157
softwar	0.01155
peopl	0.01051
compani	0.00991
comput	0.00986
year	0.00982
make	0.00963
technolog	0.00927
system	0.00896
site	0.00861
Date	**Probability**
20060512	0.10077
20060505	0.09032
20060513	0.07942
20060519	0.07665
20060524	0.07130
20060522	0.05263
20060504	0.04979
20060502	0.04776
20060523	0.04776
20060521	0.04275

analyzed. In addition, the system be altered to detect a higher number of topics; thus, increasing the granularity of cyber threat analysis.

12.4.3 Blog Content Visualization

In order to prepare the dataset, we first created a normalized 6169 × 2102 term-document matrix with term frequency (TF) local term weighting and inverse document frequency (IDF) global term weighting, based on the content of the blogs. From this matrix, we created the 2102 × 2102 document-document cosine similarity matrix, and used this as input to the dimensionality reduction algorithms. The results can be seen in Figure 12.2.

This plot is useful to see the similarities between the blog documents, and can be augmented with metadata such as blog tags or categories to visualize the distinction among blog tags. As our dataset does not contain the tags or labels, the plot is not able to show the distinction of the tags as yet.

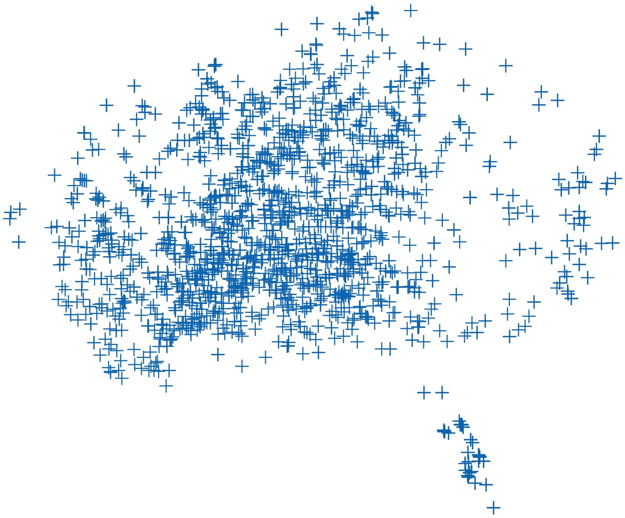

Fig. 12.2 Results on visualization of blog content using Isomap (k=12).

12.4.4 Blog Time Visualization

The dataset contained blogs from May 1-24, 2006. The date-document matrix, along with the term-document matrix, were used to compute the date-topic model. In this model, each date is represented by a probability distribution over topics, and each topic is represented as a probability distribution over terms for that topic. The topic-term and date-topic distributions were then learned from the blog data in an unsupervised manner.

For visualizing the date similarities, the symmetrized Kullback Leibler distance between topic distributions was calculated for each date pair. Figure 12.3 shows the 2D plot of the date distributions based on the date-topic distributions. In the plot, the dates were scaled according to the number of blogs in that date. The distances between the dates are proportional to the similarity between dates, based on the topic distributions of the blogs that were posted.

Viewing the date similarities in this way can complement existing analysis such as time-series analysis to provide a more complete picture of the blog time evolution.

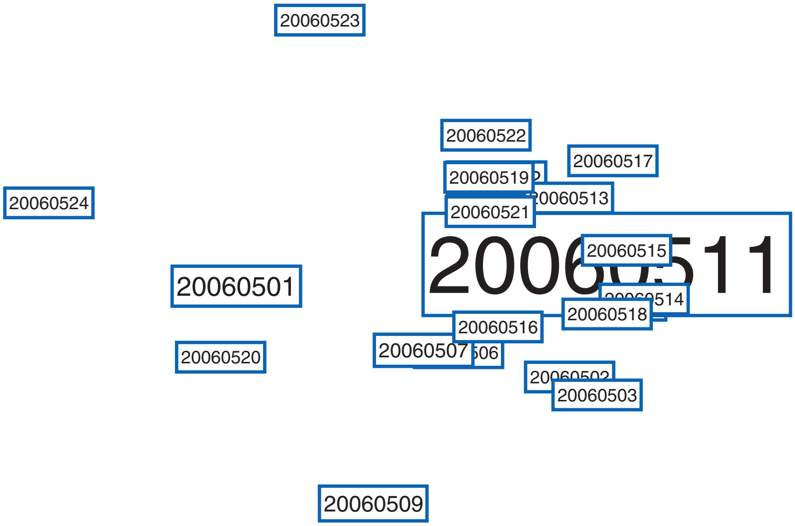

Fig. 12.3 Results on visualization of date similarities using Isomap.

They can also be further subdivided by topic to form a better understanding of topic-time relationships.

12.5 Conclusions

The rapid proliferation of blogs in recent years presents a vast new medium in which to analyze and detect potential cyber security threats in the blogosphere. In this article, we proposed blog data mining techniques for analyzing blog posts for various categories of cyber threats related to the detection of security threats, cyber crime, and information security. The important contribution of this article is the use of probabilistic and dimensionality reduction techniques for identifying and visualizing patterns of similarities in keywords and dates distributed across all the documents in the our dataset of security-related blogs. These techniques can aid the investigative processes to understand and respond to critical cyber security events and threats. Other research contributions include a proposed probabilistic model for topic detection in blogs and the demonstration of our methods for detecting cyber threats in security blogs.

Our experiments on our dataset of blogs demonstrate how our probabilistic blog model can present the blogosphere in terms of topics with measurable keywords, hence tracking popular conversations and topics in the blogosphere. By using probabilistic models, we can improve information mining in blog keywords detection, and provide an analytical foundation for the future of security analysis of blogs.

Future applications of this stream of research may include automatically monitoring and identifying trends in cyber security threats that are present in blogs. The system should be able to achieve real-time detection of potential cyber threats by updating the analysis upon the posting of new blog entries. This can be achieved by applying techniques such as folding-in for automatic updating of new blog documents without recomputing the entire matrix. Thus, the resulting system can become an important tool for government and intelligence agencies in decision making and monitoring of real-time potential international terror threats present in blog conversations and the blogosphere.

References

1. P. Avesani, M. Cova, C. Hayes, P. Massa, Learning Contextualised Weblog Topics, Proceedings of the WWW ’05 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2005.
2. D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
3. Y. Chen, F.S. Tsai, K.L. Chan, Machine Learning Techniques for Business Blog Search and Mining, *Expert Systems With Applications* 35(3), pp 581-590, 2008.
4. T. Cox and M. Cox, *Multidimensional Scaling*. Second Edition, New York: Chapman & Hall, 2001.
5. S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society of Information Science 41(6) (1990) 391–407.
6. K.E. Gill, How Can We Measure the Influence of the Blogosphere? Proceedings of the WWW ’04 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.
7. N.S. Glance, M. Hurst, T. Tomokiyo, BlogPulse: Automated Trend Discovery for Weblogs, Proceedings of the WWW ’04 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.
8. D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information Diffusion Through Blogspace, Proceedings of the WWW ’04 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.
9. M. Hickins, Congress Lights Fire Under Vote Systems Agency, Business, `www.internetnews.com/bus-news/article.php/3655001`, 2007.
10. D.H. Johnson and S. Sinanovic, “Symmetrizing the Kullback-Leibler distance,” *Technical Report, Rice University.* , 2001.
11. T. Hofmann, Unsupervised Learning by Probabilistic Latent Semantic Analysis, Machine Learning Journal 42(1) (2001) 177–196.
12. Q. Mei, C. Liu, H. Su, C. Zhai, A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs, Proceedings of the WWW ’06 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006.