BLOG ANTENNA: SUMMARIZATION OF PERSONAL BLOG TEMPORAL DYNAMICS BASED ON SELF-SIMILARITY FACTORIZATION

Yu-Ru Lin *Hari Sundaram*
Arts Media Engineering, Arizona State University
Tempe, AZ 85281
Email: {yu-ru.lin, hari.sundaram}@asu.edu

ABSTRACT

In this paper, we present a framework to analyze and summarize the temporal dynamics within personal blogs. Blog temporal dynamics are difficult to capture using a few class descriptors. Our approach comprises (1) a representation of blog dynamics using self-similarity matrices, (2) theme extraction using non-negative self-similarity matrix factorization, and (3) a visualization representing blog theme evolution. Summaries based on large real-world blog datasets reveals interesting temporal characteristics for four blog types – personal blog, cooperative blog, power blog and spam blogs.

1 INTRODUCTION

In this paper, we propose a new framework to summarize *personal* blog activity. Blogs represent highly popular, new media for communication. Summarization is important as these media are characterized by the content temporal dynamics as well as high posting volume – these are difficult to capture using static tag clouds or time-based tag streams [4], or a few class descriptors. An example of the proposed summary is (where arrows are posts, and the colors represent the themes) shown in Figure 1.

Figure 1: Blog Summary

There has been prior work on understanding the temporal dynamics of the *entire* blogosphere. Chi et al. [1] proposes "eigen-trend" to represent the temporal trend in a group of blogs with common interests by using singular value decomposition (SVD). Mei et al. [6] consider a problem of mining spatiotemporal theme patterns from blogs based on a probabilistic approach and the "theme snapshots" representation. In this research we focus on analyzing content themes (rather than explicit keywords) in a single personal blog, not the entire blogosphere.

In multimedia, there has been prior work on temporal structural analysis of audio signals. In [5] the authors propose self-similarities as a way to visualize musical structures. They have applied SVD and non-symmetric NMF to media segmentation or summarization [2]. In this work we generalize this research to arbitrary non-numeric time series data (blog-posts) as opposed to analyzing continuous auditory data.

There are three key ideas in our analytical framework to discover blog temporal dynamics. We first represent the blog temporal sequence using self-similarity matrices defined on the histogram intersection similarity measure of the content and link attributes of posts. Second, the temporal relationship of posts is determined as clustering using symmetric non-negative matrix factorization of the self-similarity matrices, and the clustering quality is determined by the "modularity function". Finally, we summarize the blog temporal dynamics using a "blog antenna" summary based on the similarity factorization results. The blog antenna simultaneously reveals the relationship (similarity of content / link) of blog post sequence with respect to time. The summary also shows a projection view to reveal theme evolution and time independent theme relationships. We have tested the summarization on four types of blogs from the TREC dataset – personal, cooperative (community blog), power blog and spam blog (splog) – with excellent results.

The rest of the paper is organized as follows. In next section we examine blog topological matrices. We present our method of determining and representing the temporal relationship of blog post sequence in section 3 and section 4. We show results of experimental studies in section 5 and then present conclusions.

2 SELF-SIMILARITY

We now analyze the temporal dynamics of blog by examining the temporal self-similarity of its prime attributes, such as the post content, citation links, tags, etc. The intuition behind using self-similarity lies in its ability to reveal temporal structures.

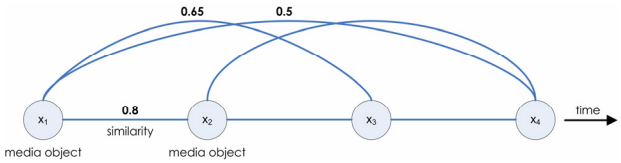

Figure 2: A topological graph induced on time sequence of media objects (e.g. blog posts) x_i, due to a specific similarity measure $s(i,j; a_k)$ on attribute a_k.

Let us assume that we have a sequence of N media objects (e.g. blog posts) x_i, from the same blog that are ordered in time – i.e. $t(x_i) \le t(x_{i+m})$ for all $m \ge 0$. Assume that the media objects can be described using k attributes, and that for each attribute a_k, we can define a similarity measure $s(i,i+m; a_k)$. This measures the similarity between objects x_i and x_{i+m}, using attribute a_k. This measure induces a point set topology over the set x_i. We can create an attribute dependent topological matrix S_k using the topology where the elements of the matrix $S_k(i,j; a_k)$ are defined as follows: $S_k(i,j; a_k) = s(i,j; a_k)$, $i,j \in \{1,\ldots, N\}$. Since the media objects are ordered in time, the topological matrix reveals the temporal self-similarity of the sequence x_i with respect to attribute a_k. Figure 2 shows an example topological graph.

The topological matrix functions as a generalized autocorrelation over any time series of media objects—if we average all the values along each diagonal in the upper triangular matrix, this would be equivalent to computing the autocorrelation of the non-numeric sequence. It is important to keep in mind that the nodes in Figure 2 refer to *posts from the*

same blog and the edges refer to the similarity between posts along a specific attribute. We now examine the topological matrixes computed on the temporal sequences of blog posts. We focus on two prime attributes: post content and links.

Post content. The similarity measure to induce the topology on post content is defined using histogram intersection similarity measure on the tf-idf vectors. Let h_i and h_j be the tf-idf vectors (after stemming and stop-word removal) for posts p_j and p_j. Then the similarity between the posts is defined as:

$$S(p_i, p_j; c) = \sum_{k=1}^{M} \frac{\min\left(h_i(k), h_j(k)\right)}{\max\left(h_i(k), h_j(k)\right)} \quad <1>$$

where, c refers to the similarity based on the content attribute and M is the size of the vector. Note that the posts are ordered in time. The corresponding element in topological matrix is then the content similarity between the two posts.

Figure 3a shows an example plot of the content-based temporal self-similarity topological matrix. It reveals that there is significant temporal correlation between posts. It also shows that the users mostly tend to stay on a topic (large white blocks in the figure), and may occasionally post on different topics (causing black horizontal lines) before returning to the topics.

Post links. The similarity measure on the links is defined in a similar manner to eq. <1> except that the tf-idf vectors are now calculated on the target links (stemmed by the root domain), rather than on the words. Hence:

$$S(p_i, p_j; l) = \sum_{k=1}^{P} \frac{\min\left(h_i(k), h_j(k)\right)}{\max\left(h_i(k), h_j(k)\right)} \quad <2>$$

where, l refers to the similarity on the link attribute, h_i and h_j are the link based tf-idf vectors and P is the size of the vector.

In Figure 3b, we can see the link based self-similarity matrix – it reveals a similar block based structure and changes as in Figure 3a, and we can see that changes to the content and link patterns are usually coincident.

The self-similarity matrix exhibits several advantages: First, the media objects are represented by their relationship with the other objects. The size of the matrix used to represent a dataset of N objects is always $N \times N$, regardless the content complexity of the objects. Second, as the self-similarity matrix is constructed according to the time order of objects, it allows investigation of the temporal relationship among objects.

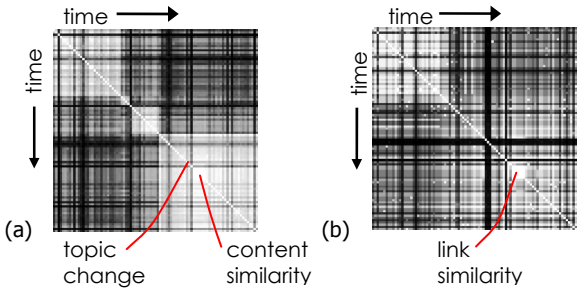

Figure 3: (a) Self-similarity matrix on the content attribute. (b) Self-similarity matrix on the link attribute.

3 FACTORIZATION OF THE BLOG TEMPORAL DYNAMICS

We shall now discuss our method for summarizing the blog temporal dynamics. The content of a typical human created blog usually exhibits short-term coherence with topic drifting over a long period. To summarize the temporal dynamics, let us first examine the relationship between two posts. Intuitively, if two posts are similar in terms of attribute values, it is possible that these two posts are highly related to certain common things that might not be explicitly stated in the content—we refer the common things as the underlying "themes". For example, two posts that have terms such as "Star Treck" in common are likely to have a sci-fi or TV-series theme. Naturally, a post can relate to multiple themes. We conjecture the similarity relationship of two posts reflects how their underlying themes coincide, and by using a few common themes and the relationship between posts and themes we are able to summarize the temporal variation of posts.

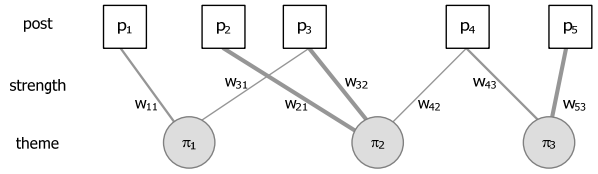

Figure 4: The post-post relationship is summarized by post-theme relationship.

Let us assume each post is governed by a set of underlying themes with different strengths, as illustrated in Figure 4. Let w_{ic} be the strength that a post p_i is governed by a theme π_c. We seek to estimate w based on the observed similarity relationship of posts. Assume there are N posts and K underlying themes, the objective function can be expressed as:

$$\arg\min \sum_{i=1}^{N}\sum_{j=1}^{N} \left\| s_{ij} - \sum_{c=1}^{K} w_{ic} w_{jc} \right\| \quad <3>$$

where s_{ij} is the similarity of two posts p_i and p_j. We can rewrite eq.<3> as:

$$W^* = \arg\min_{W} \left\| S - WW^T \right\| \quad <4>$$

where S is an $N \times N$ matrix with element $[S]_{ij}=s_{ij}$ and W is an $N \times K$ matrix with element $[W]_{ic}=w_{ic}$. The solution to eq.<4> can be obtained by symmetric non-negative matrix factorization (NMF), where the solution W^* is approximated by iterative updating as described in [3].

The rank K of the factorization is chosen using a clustering quality criterion: the "modularity function" [7]. Here the modularity function Q is defined as:

$$Q(k) = \sum_{c=1}^{k} \left[\frac{\sum_{i,j} s_{ij} w_{ic} w_{jc}}{\sum_{i,j} s_{ij}} - \left(\frac{\sum_{j} s_{ij} w_{ic}}{\sum_{i,j} s_{ij}} \right)^2 \right] \quad <5>$$

where k is the rank used in the factorization and K is selected so as to maximize $Q(k)$, i.e. $K=\text{argmax}_k\ Q(k)$. Consider a simple case where exactly one element in each row is set to 1 and the rest are set to 0, i.e. $w_{ic}=1$ indicates whether post p_i belongs to theme π_c. The first term in the bracket of right hand side in eq.<5> expresses the empirical joint probability pr[c,c] that two

posts belong to the same theme π_c, and the second term is the empirical probability $\text{pr}[c]^2$ and $\text{pr}[c]$ is the probability of a post belonging to π_c. If the estimation of w is no better than random, we will get Q=0. In general, a larger value of Q indicates the estimation of w better explain the empirical probabilities.

The matrix S describes the similarity relationship of posts. We can derive S from self-similarity matrices on multiple attributes. S is computed as follows:

$$[S]_{ij} = s_{ij} = \sum_k \alpha_k s(i,j;a_k) \quad <6>$$

where $s(i,j; a_k)$ refers to similarity measure on attribute a_k, and α_k refers to the weight on the corresponding attribute. Here we consider the similarity relationship as a linear combination of different similarity measures in post content and links. The dominant theme of a post is determined as follows:

$$c_i^* = \arg\max_{1 \le l \le K} w_{il} \quad <7>$$

where c_i^* is the dominant theme indicator of post p_i.

4 SUMMARIZATION

We now discuss our proposed visual summary for blog temporal dynamics. Our goal is to have the summary representation answer the following questions – (1) How many distinct or major themes exist in the blog? (2) how do the sequence of posts correspond to these themes? (3) reveal how the themes evolve over time? (4) is there a relationship among themes independent of time? and finally, (5) can we observe a blog from all the above aspects?

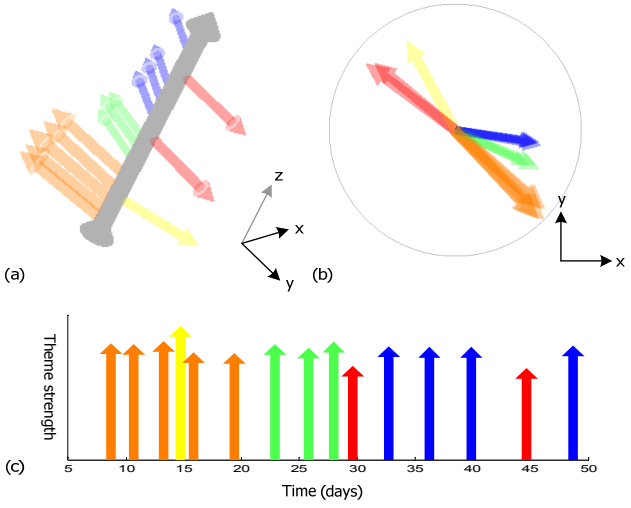

Figure 5: Exploring blog temporal dynamics from three perspectives: (a) 3D antenna view: Each post is rendered around the time axis, with position representing the post time, length representing the dominant theme strength and orientation representing the content similarity comparing to other posts. (b) Content drifting view: The orientation of post-arrows represents the similarity between the content of posts, and length represents the dominant theme strength. (c) Content evolving view: Each arrow represents a post, stemming from the position where the post is created. Each post-arrow is colored by its dominant theme, with length representing the dominant theme strength.

Our visualization uses a metaphor "antenna" to represent how blog content evolves over time. As shown in Figure 5, we render each post as an arrow-shaped antenna stemmed from a time axis. The older the posts, the closer they are to the origin of the time axis. By default, posts are colored with respect to their dominant themes, with different colors for different themes. The length of an arrow indicates the strength of the post corresponding to the dominant theme.

The visualization provides three perspectives—the 3D antenna view, the content drift view and the content evolving view. We discuss each in detail below.

1. *3D antenna view* (Figure 5a): The view immediately reveals post diversity (colored themes, different directions), and post density. We represent the post, themes, and their relationships with respect to time simultaneously using a three dimensional antenna object. The object comprises a central time axis along the z-axis direction, and each post is rendered around the time axis, with position representing the post time, length representing the dominant theme strength and orientation representing the content similarity comparing to other posts.
2. *Content drifting view* (Figure 5b): This view reveals time-independent theme dissimilarity. We seek to render similar posts as arrows which ends are close in distance, while dissimilar posts are rendered far apart. This is computed by a multiple dimensional scaling technique based on principle component analysis. Specifically, we reconstruct a matrix S' from the NMF factors W by $S'=WW^T$, where W is obtained from eq.<4>. Let $S'=U\Sigma V^T$ be the singular vector decomposition of S', where U and V are $n\times k$ and $m\times k$ orthogonal matrices and Σ is a diagonal matrix whose diagonal elements are the singular values of S'. Because S' is symmetric, we have $U=V$ and $n=m$. Here we use k=2 because a two-dimensional representation allows effective comparison of the relationship among themes and posts, based on the arrow orientation and length. We then map the two principal components $v_i=[V]_{i,1}$ and $u_i=[V]_{i,2}$ of each post p_i on to a 2D x-y plane, as shown in Figure 5b. The post-arrows of opposite directions, such as the blue and red arrows, have the most dissimilar content.
3. *Content evolving view* (Figure 5c): This view is useful for comparing posts in terms of the strength with respect to dominant themes. The time axis lies at bottom and all post-arrows point upward. In an alternative representation (Figure 6), posts are shown as uniform length arrows with multiple segments, each segment represents a theme, and the length of the segment indicates the relative strength of corresponding theme, in that post.

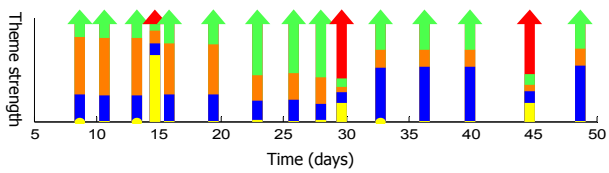

Figure 6: An alternative representation of content evolving view. Each segment of a post represents a theme, with the length of the segment representing the relative strength of the post to the corresponding theme.

5 EXPERIMENTS

We now present summaries of blogs using blog antenna representation based on our preliminary experiments. We demonstrate the effectiveness of the analytical framework for discovery of blog temporal characteristics. We use the TREC (the Text Retrieval Conference) Blog-Track 2006 dataset for analysis. This dataset is a crawl of 100,649 feeds collected over 11 weeks, from Dec. 6, 2005 to Feb. 21, 2006, totaling 77 days. We focus our analysis on a subset of English blogs having homepage and at least one entry.

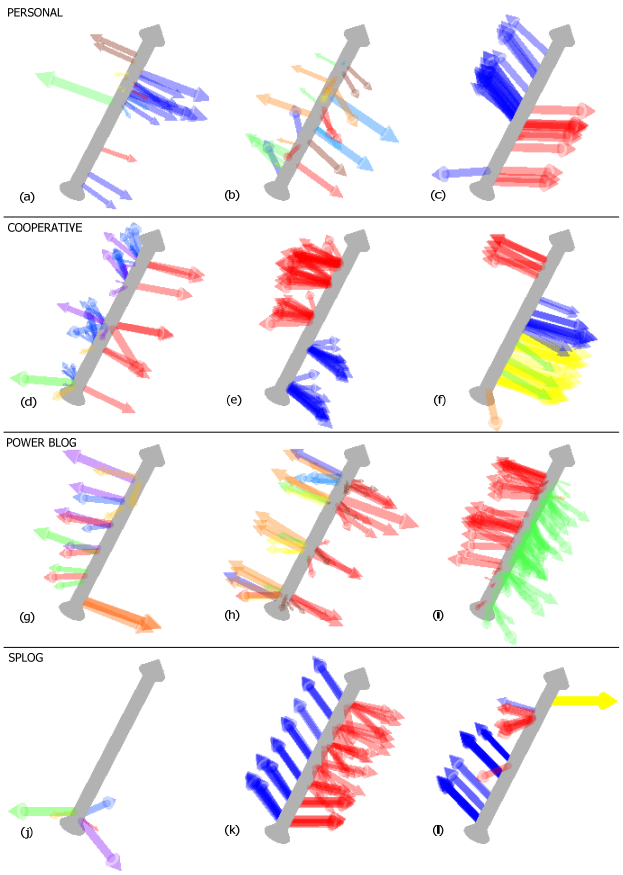

Figure 7: Examples of blogging antennas. (a,b,c) Personal blogs; (d,e,f) Cooperative blogs; (g,h,i) Power / content aggregator blogs; (j,k,l) splogs.

Examples of blog antennae. We have observed interesting temporal characteristics of different types of blogs from the experimental studies. The following are representative examples:

1. *Personal blog* (Figure 7a,b,c): Personal blogs have diverse content posted at irregular times, the posts are usually spread out in time, and the post intervals are larger than other blog types. Figure 7a shows a blog with one major theme where posts of other themes are interleaving. Figure 7b shows another blog with multiple themes in the blog. The blogger in Figure 7c dedicates to one topic for some time and changes to another afterward.
2. *Cooperative blog* (Figure 7d,e,f): This type of blog has more than one author posting or discussing things of their common interests. Cooperative blogs can be characterized by the short-term bursty phenomenon – typically an external event triggers a succession of posts. Figure 7d,e,f show three cooperative blogs that discuss politics, local news and information, and design / art of advertising, respectively. In the last blog, in spite of multiple authors, there is a dominant blogger who keeps posting, which makes it close to a personal blog.
3. *Power blog / content aggregator* (Figure 7g,h,i): This type of blogs are usually maintained by dedicated bloggers who regularly provide news or reviews related to specific topics. Such blogs can be characterized by periodic or frequent posting activities with concentrated (sometimes alternating) topics. For example, Figure 7g shows a blog aggregating news about Mac in a weekly basis. Figure 7h shows another blog reporting oil prices and related news in consecutive posts simultaneously. Figure 7i is a professional blogger actively posts his idea about innovation.
4. *Splog* (Figure 7 j,k,l): The spam blogs (or splogs) have very distinct patterns from human blogs. It can be seen that these splogs either "die" very soon (Figure 7j), or have posts created in a precisely regular basis, with almost identical content (Figure 7 k,l).

6 CONCLUSION

In this paper we propose an analytical framework and the blog-antenna visual metaphor to summarize personal blog temporal dynamics. Our approach comprises three steps: (1) a representation of temporal sequence of blog content using self-similarity matrices, (2) temporal content variation analysis based on non-negative self-similarity matrix factorization, and (3) a visual representation of the blog temporal dynamics. Our experiments on TREC data reveals characteristic blog antennas for blogs such as personal, cooperative, power blogs and splogs. We plan to extend our work to address normalized timescales as well as a multi-scale knowledge summary.

7 REFERENCES

[1] Y. CHI, B. L. TSENG and J. I. TATEMURA (2006). *Eigen-trend: trend analysis in the blogosphere based on singular value decompositions*, CIKM 2006, 68-77,

[2] M. COOPER and J. FOOTE (2002). *Summarizing video using non-negative similarity matrix factorization.* Multimedia Signal Processing, 2002 IEEE Workshop on: 25-28.

[3] C. DING, X. HE and H. SIMON (2005). *On the equivalence of nonnegative matrix factorization and spectral clustering.* Proc. SIAM Data Mining Conf.

[4] M. DUBINKO, R. KUMAR, J. MAGNANI, J. NOVAK, P. RAGHAVAN and A. TOMKINS (2006). *Visualizing Tags over Time*, International World Wide Web Conference, ACM, 193-202, 2006///.

[5] J. FOOTE, M. COOPER and U. NAM (2002). *Audio retrieval by rhythmic similarity.* Proceedings of the International Conference on Music Information Retrieval **3**: 265–266.

[6] Q. MEI, C. LIU, H. SU and C. ZHAI (2006). *A probabilistic approach to spatiotemporal theme pattern mining on weblogs.* Proceedings of the 15th international conference on World Wide Web: 533-542.

[7] M. NEWMAN and M. GIRVAN (2004). *Finding and evaluating community structure in networks.* Physical Review E **69**(2): 26113.