Socio-temporal analysis of conversation themes in blogs by tensor factorization

Abstract

Blogs have been popular on the Internet for a number of years and are becoming increasingly popular within the organizations as well. The analysis of blog posts, especially in the organizational context, is a way to understand the *topics* of *conversations* that evolve through the organization. While keywords within blog posts can be used to characterize the topic being discussed in a post, the timestamps permits one to distinguish among the objects of the discussion, and the authors of posts provide a mean of separating different perspectives on the matter. Based on this observation we define *themes of conversation* using *keywords*, *people* and *timestamps* of the posts. We represent the data as a tensor and show that higher order factorization of the tensor can separate themes of conversation, identify important people in each such theme, and determine the activity level of each theme over time based on the importance of the participants and the keywords. We evaluate this approach by applying it to a dataset extracted from a blog network within a large globally distributed IT services provider over 22 months. We discuss implications of this work for monitoring opinion developments and detecting opinion leaders within the organization.

1. Introduction

Automated analysis of large scale blog data to gather organizational intelligence is a topic of considerable interest to managers. The last decade has seen significant progress by the data mining community in blog data analysis with minimal human intervention. Especially interesting are those that exploit multi-modal data, e.g., data containing authors, relations, time, topics etc., because they attempt to understand the dynamics of the blog network without simplifying them to only the text in the posts or to only the relations between the actors (Yun, Belle et al. 2006; Shenghuo, Kai et al. 2007; Yun, Shenghuo et al. 2007). The current work builds on this literature and focuses on identification of significant themes as part of the larger investigation into mapping expertise within the organization.

Examination of the intra-organizational blog dataset collected for this purpose reveals that most of the posts are chatters: collection of jokes, quotes, and interesting news items etc. that form no consistent pattern and generate interest for only a short duration. However, buried in the chatter, there are long running topics driven by one or a small group of bloggers who could be considered authorities in the subject (See Table 1). In this paper we work with the intuition that a theme of posts can be considered significant if it is posted by authorities in the topic, has endured for a period of time, and has significant content. The needed importance scores are not available a priori. However, they can be determined from the occurrence patterns of authors and keywords over time.

Id: xxx081 Date: 2007-09-05 *Diodes can be classified by the functions of the circuit in which it is being used, or more commonly, by the shape that is demanded by the size of the products in which it will be mounted. The complicated point is that there is no direct relation between the two …* (125 more posts by xxx081 in next ten days on "voltage", "diodes" and "semiconductors")	**Id: xxx991 Date: 2007-11-09** *Benefits of Automated Testing. If you have ever tested applications or Web sites manually, you are aware of the drawbacks. Manual testing is time-consuming and tedious, requiring a heavy investment in human resources. Worst of all, time constraints often make it impossib…* (150 more posts by xxx991 in next eight weeks on "software", "test", "automation")	**Id: xxx368 Date: 2007-10-10** *20 Minute Home Work Out. If you are busy, not able to get up early morning or have no time for gym just follow this 20 minute home work out to stay healthy and fit. 1) Jog in one place for 3 minutes. Simple light jogging on the spot. 2) Jumping jacks: 25 repeats When landing …* (190 more posts by xxx368 in next hundred days on "exercise", "muscle", "weight")

Table 1 Some of the topics in a blog network along with posting pattern of people behind them.

2. Theme identification

Singular value decomposition of an adjacency matrix of a network results in the hub and authority scores of the nodes (Pagerank, HITS (Kleinberg 1999)). The leading left singular vector gives the hub scores whereas the leading right singular vector gives the authority scores. The node with high hub scores are the ones that link to nodes with high authority scores and the nodes with high authority scores are the ones linked to by nodes with high hub scores. Usually the leading singular vector pair is used since they explain most of the data, however subsequent singular vector pairs can also be used if their singular values indicate that they explain substantial portion of the data as well. Subsequent pairs have the same relation between the hubs and the authorities. Each pair corresponds to a different community sub-structure over the nodes. The first k pairs of singular vectors provide a decomposition of the two dimensional data matrix into k rank-1 matrices. This method is unsupervised: topics are determined solely from the co-occurrence patterns in the data.

However, not all datasets can be satisfactorily represented by a two dimensional matrix. In a blog network where relations are indicated by citations and replies, encoding the relation by a single number would lose the content and the context of the relation. Or, in the case of an evolving network, where there is a timestamp associated with each tie, a two dimensional representation of the relational data would have to be at the expense of temporal information. Such data is better represented and analyzed in a tensor. One example is TOPHITS that extends the HITS algorithm by associating anchor text with the hyperlinks (Kolda and Bader 2006). We present two application of tensor decomposition for blog data analysis.

Blog topic developments

One view of the blogs is that they are self-publication media where bloggers write on topics of their interest. If the goal is to identify different development themes of posts, one needs to look beyond the word occurrences in the blog posts. Spikes of posts containing same keywords in two separated time periods are often about different subjects, e.g., posts with keywords related to "hurricane" published in last week of Aug '05 are likely to be about the hurricane Katrina, whereas the posts with similar keywords in the last week of Sep '05 are likely to be about hurricane Rita. Similarly, posts containing same keywords made by two different people are

likely to differ in the perspective offered in the same topic. Therefore, to identify different themes in blog posts, the relevant variables are the authors, timestamps and keywords of the blog posts. This data can be represented as a $author \times keyword \times timestamp$ tensor $\mathcal{X}$, where, each cell of the tensor contains *tf-idf* weighted and length normalized counts of the word occurrences. This value indicates the strength of co-occurrence of the three variables. Consider the following reinforcing definition of authority of bloggers, importance of keywords and intensity of a topic at a given time period for a particular topic:

1. The authority of a blogger in a topic can be judged from her participation during the period when the intensity of the topic is high and from her use of important keywords.
2. The importance of a keyword in a topic can be judged from its use by the authorities in the topic and from its use during the period when the intensity of the topics is high.
3. The intensity of a topic during a time period can be identified from the number of posts made by authorities and the presence of important keywords in the topic.

This is a higher order extension of hub and authority scores. When there is only one topic, according to the definition, the importance of the bloggers in this topic can be calculated as:

$$a_p = \sum_q \sum_r x_{pqr} k_q t_r \Leftrightarrow \boldsymbol{a} = \mathcal{X} \times_2 \boldsymbol{k} \times_3 \boldsymbol{t}, \text{similarly } \boldsymbol{k} = \mathcal{X} \times_1 \boldsymbol{a} \times_3 \boldsymbol{t} \text{ and } \boldsymbol{t} = \mathcal{X} \times_1 \boldsymbol{a} \times_2 \boldsymbol{k}$$

where, $\mathcal{X} \in \mathbb{R}^{|\boldsymbol{a}| \times |\boldsymbol{k}| \times |\boldsymbol{t}|}$; $\boldsymbol{a}, \boldsymbol{k},$ and $\boldsymbol{t}$ are the vectors of importance of the authors, keywords and the time periods. $\times_j$ is the j-mode product of a vector with a tensor. Applied iteratively the vectors $\boldsymbol{a}, \boldsymbol{k},$ and $\boldsymbol{t}$ converge to minimize the error $\|\mathcal{X} - \boldsymbol{a} \circ \boldsymbol{k} \circ \boldsymbol{t}\|_F$, where, $\circ$ is the outer product between the vectors (De Lathauwer, De Moor et al. 2000). Thus $\boldsymbol{a} \circ \boldsymbol{k} \circ \boldsymbol{t}$ is the best rank-1 approximation of the tensor $\mathcal{X}$. Extending to R topics and using a normalizer λ to make each vector of unit length the decomposition can be expressed as sum of R rank-1 tensors:

$$\mathcal{X} = \sum_r^R \lambda_r \times \boldsymbol{a_r} \circ \boldsymbol{k_r} \circ \boldsymbol{t_r} = [\boldsymbol{\lambda}; \mathbf{A}, \mathbf{K}, \mathbf{T}]$$

where, $\mathbf{A}, \mathbf{K},$ and $\mathbf{T}$ are three matrices with R $\boldsymbol{a_r}, \boldsymbol{k_r},$ and $\boldsymbol{t_r}$ vectors as columns respectively. This decomposition is known as Parallel Factorization (PARAFAC) (Harshman 1970). This decomposition can be computed by Alternating-Least-Square error minimization (ALS). The error $\|\mathcal{X} - [\boldsymbol{\lambda}; \mathbf{A}, \mathbf{K}, \mathbf{T}]\|_F$ is minimized by successively optimizing one of the three matrices while keeping the other two constant. The detailed ALS algorithm can be found in (Kolda and Bader 2008). An implementation is available in their TensorToolbox matlab package (Bader and Kolda 2007).

Blog conversation development

In this extension we take into account the conversational nature of the blog posts. A comment to a blog post or a post with a citation has an author and a recipient. Content of the post not only

depend on who is making the post but also who is it targeted to. To capture this fact we represent the blog data in a fourth order tensor ($author \times recipient \times kewords \times timestamp$). The idea behind evaluating the importance of a variable is similar to that in blog topic development analysis. The extension is that the importance of the recipient of the conversation influences the importance of the variables in other modes.

3. Experiments and Results

Data description and representation

The data for this study is collected from a private blog network in a large IT services firm. It contains the blog posts and replies along with timestamps and demographic information about the bloggers. We used two subsets of the data for the methods described in Section 2. For blog topic development analysis we used the text of the blog posts and the comments, author-ids, and the timestamp. For the blog conversation development analysis we used only the comments on the blog post. We excluded the blog posts because it is not targeted to anyone. Each comment can be considered to be targeted to the original poster. Therefore, the original poster-id is used as the recipient. Thus the variables used are the text in the comments, the commentor-id, the recipient-id, and the timestamp.

Date range	May '06—Jan '08		
Blog posts	51,000 average length 300 words		
Comments	131,000 average length 33 words		
Bloggers	3600	Commentors	14,000
Blogs	2950		

Results and discussion

Each tensor was decomposed into 20 rank-1 tensors using PARAFAC. Some of the resulting factors are displayed in Figure 1 and Figure 2. For each factor the top five keywords are shown along with the importance of the top-authors and the daily intensity of the topic.

We manually examined the posts made by the top-authors and compared them with those made by the other authors. As an example, in the "Indian political history" topic, the lead author's first few posts were chatters. They were ignored by the algorithm. After that the blogger posted a series of articles that were focused on the subject. The decomposition detected them as a significant theme and the blogger as an authority in the subject. In general the most significant author in a topic was easily identifiable as being an expert in the area. However, the second and third significant authors were found to be only loosely related to the topic.

The intensities of the topics over time tell us the nature of the conversation: posts about "software testing" have generated interest for much shorter period compared to the conversations about "new technology companies". The importance scores of the authors also give insight into the nature of the conversation. Posts about physical exercise have seen activity over about 100 days, but, they are primarily made by one person. On the other hand more people have contributed to "software testing" and "Indian political history" topics, though they were active for a shorter period.

Analysis of comments on the blog posts reveals a set of factors that generate higher than average amount of reactions from the bloggers, e.g., religion, mythology and law. The effectiveness of the decomposition is illustrated by the latent semantic separation of *religion & food*, *spirituality* and *mythology* that can be thought of as a part of broader topic of religion. Though there are more authors who can be considered important in this analysis, we do not observe multiple important recipients in any topic due to the nature of the conversation in the blogs. There are certain blogs that are popular places for discussing certain topics. In a sense these blogs are the harbors for conversations in the topic and they play a central role in defining the ongoing conversations as a topic.

4. Conclusion

We present an approach to identify important themes of conversation in a blog network. The highlight of the approach is the simultaneous determination of the importance of a theme and that of the *text*, *author* and *timestamp* that define the theme by one simple process of tensor decomposition. Applying to a fourth order tensor that has *author* as well as the *recipient*, we extend the analysis into the domain of conversations. This can also be seen as a generalization of the network centrality computation in that the link between two individuals in not only a number, but, a text document expressed as a vector of term weights. We shows this allows finer grained importance computation of authors and recipients.

5. Bibliography

Bader, B. W. and T. G. Kolda (2007). MATLAB Tensor Toolbox, version 2.2.

De Lathauwer, L., B. De Moor, et al. (2000). "On the Best Rank-1 and Rank-(R, R,..., R) Approximation of Higher-Order Tensors." SIAM Journal on Matrix Analysis and Applications **21**: 1324.

Harshman, R. A. (1970). "Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis." UCLA Working Papers in Phonetics **16**: 1-84.

Kleinberg, J. M. (1999). "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) **46**(5): 604-632.

Kolda, T. G. and B. B. Bader (2008). "Tensor Decompositions and Applications." SIAM Review.

Kolda, T. G. and B. W. Bader (2006). "The TOPHITS model for higher-order web link analysis." Workshop on Link Analysis, Counterterrorism and Security.

Shenghuo, Z., Y. Kai, et al. (2007). Combining content and link for classification using matrix factorization. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Amsterdam, The Netherlands, ACM.

Yun, C., L. T. Belle, et al. (2006). Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. Proceedings of the 15th ACM international conference on Information and knowledge management. Arlington, Virginia, USA, ACM.

Yun, C., Z. Shenghuo, et al. (2007). Structural and temporal analysis of the blogosphere through community factorization. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. San Jose, California, USA, ACM.

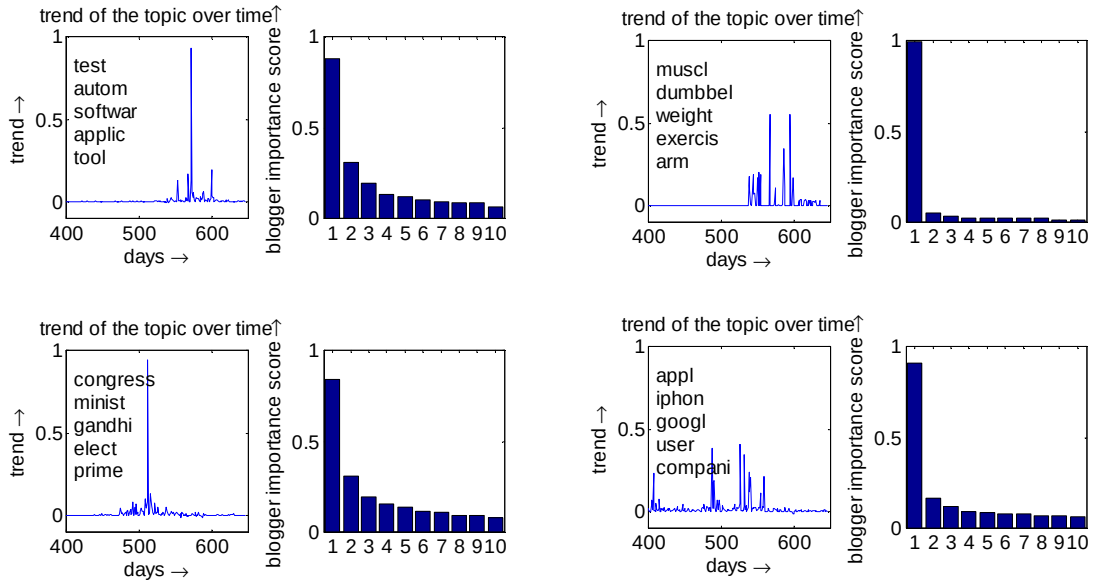

Figure 1 Trends of topics and important bloggers in each.

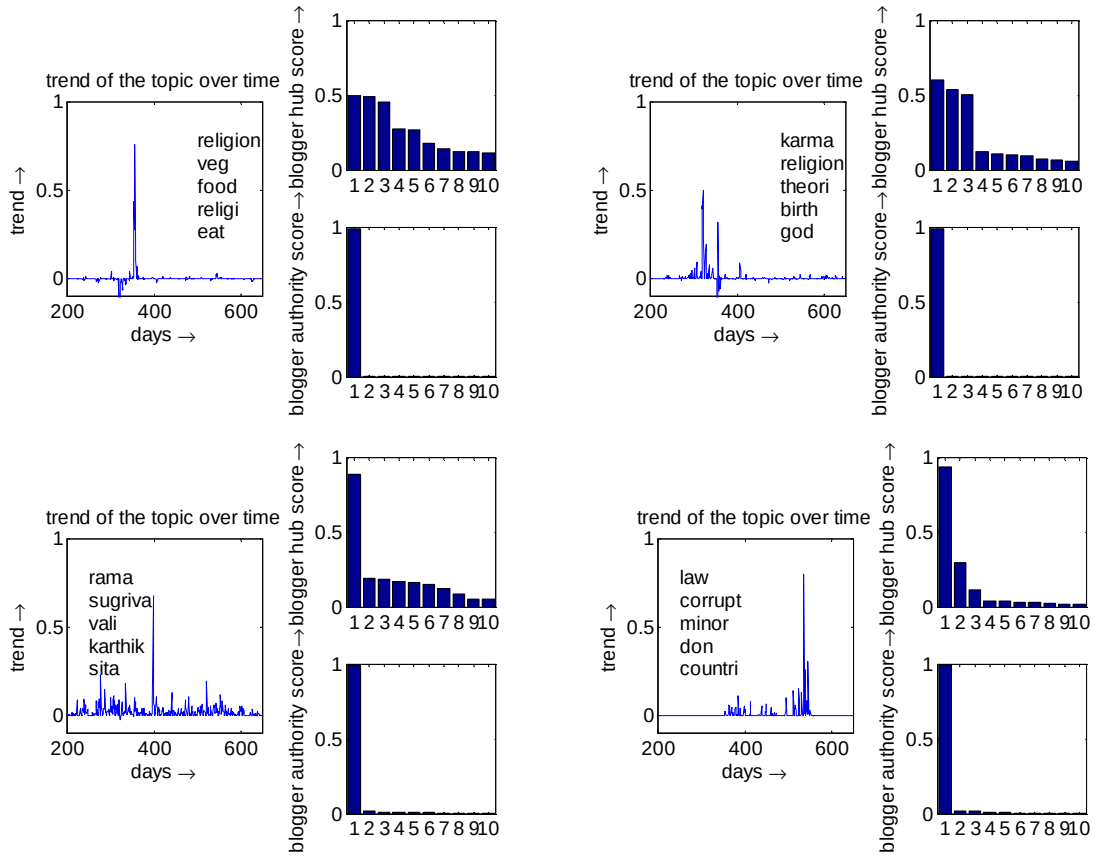

Figure 2 Trends of topics and topic specific hub and authority score