Social Media Analysis via Network Approaches

Victor Cheung, Zhi-Li Wu, Chun-hung Li
Department of Computer Science
Hong Kong Baptist University
{vincent,victor,chli}@comp.hkbu.edu.hk

Abstract

Social media such as online forum and weblog are composed of dense interactions between user and content where network models are often appropriate for analysis. Using Markov logic network, user participation models can be developed to help us gain insights on the latent base topics of online discussions. Furthermore, joint non-negative matrix factorization model of participation and content data which can be viewed as a bipartite graph model between users and media. The factorizations allows simultaneous automatic discovery of leaders and sub-communities in the online forum as well as the core latent topics in the forum. Results on topic detection of online forums as well as the clustering analysis are given. Part of this work is based on a copyright-free paper that received Best Workshop Paper Award in 2007 IEEE International Conference on E-Business Engineering Student Workshop.

1. Introduction to Online Forum

Recent years have seen a greatly increased attention to online communities where internet users play a dominant role in content contribution and sharing. Numerous research has been conducted to understand the motivations of users in online community [4] and the method to increase participation [1]. Social network analysis [2] has also been used to analyze online forum. Nolker and Zhou apply social network theory to newsgroup to discover leaders, motivators and chatters [6].

In an online forum, usually one user will post a topic which can be a question or a piece of information. The Zipf's law stated that the frequency of any English word is inversely proportional to its rank in the frequency table [10]. Figure 1 shows the log-log plot of frequency of user participation versus that of the rank of that user in an online forum. From the plot we can see that the user participation frequency follows the Zipf's law. In other words, active users of the online forum can post significantly more than other users. There are lots of users who post only rarely. This uneven posting frequency distribution pattern motivates the use of special normalization method in the processing of user participation data.

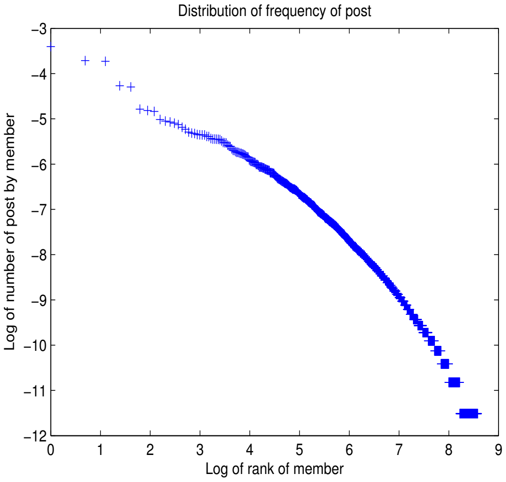

Figure 1. Distribution of Post Frequency vs User Rank

2 User participation for clustering via Markov Logic Networks

It is possible to demonstrate the grouping effect of discussions as reflected from user participation. A discussion-discussion similarity can be formed from markov logic networks (MLN) [8], which evaluates discussion similarity by specifying logic formula to explore relationships among discussions and users. Traditional first-order Knowledge Base can be regarded as a set of hard constraints on possible worlds. If a world violates even one of them, the word has zero probability to exist. For markov logic networks, there are also a set of constraints, expressed in formula of first order logic. However, constraints in MLNs are softer: when a world violate any of them, the existence of the world is less probable, but not impossible. Each of the formula is associated with a scalar weight reflecting the importance or hardness of the formula. The probability of existence of a world depends on the weighted sum of the satisfied constraints. The world is more probable if the sum is higher.

In this paper, MLN is applied to topic detection of a forum and it is assumed all formulas are function-free clauses and domain closure. Inference with MLNs can be done by inferring on the grounded Markov network. One of the most widely used approximate algorithms to the evaluation is Markov chain Monte Carlo (MCMC), and in particular Gibbs sampling, which proceeds by sampling each variable in turn given its Markov blanket, and counting the fraction of samples that each variable in each state.

After the similarity matrix is evaluated, a public domain available clustering tool CLUTO is employed to cluster the discussions. A transformed view of the clustering solutions in 3-dimension is shown in Figure 2 based on MLN similarity.

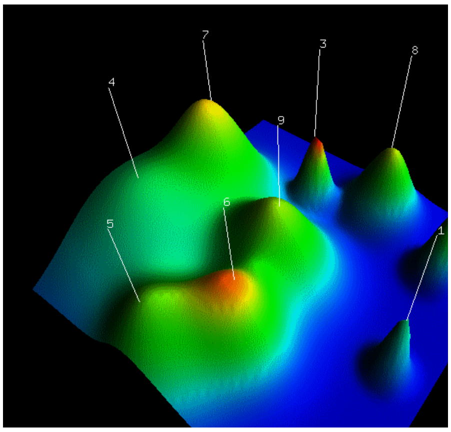

Figure 2. Transformed 3-D view of the clustering solution for the similarity matrix formulated by using MLN.

3. Non-negative Matrix Factorization

Nonnegative matrix factorization (NMF) has been studied under the name of positive matrix factorization [7] at the early stage. It is then popularized by the work of Lee and Seung and has been found lots of applications in text mining [5],[9]. Recent advancement of NMF has shown that it shares much similarity with K-means and spectral clustering methods, and is capable of producing good cluster capability [3].

In the following, we briefly review and extend NMF to factorize two closely related non-negative matrices. Given a non-negative matrix X in size $n \times m$, NMF factorizes it into two nonnegative matrices W and H,

$$X = WH,$$

where W is a $n \times k$ matrix and H is $k \times m$, while k is usually much smaller than both n and m.

To minimize the square Euclidean distance between X and the reconstructed matrix $\tilde{X} = WH$,

$$\min_{W\geq 0, H\geq 0} ||X - \tilde{X}||^2 = \sum_{i=1}^{i=n}\sum_{j=1}^{j=m}(X_{ij} - \tilde{X}_{ij})^2$$

The objective function can be iteratively reduced, or be kept non-increasing, via the following updating rules,

$$H = H. * (W^T X)./(W^T WH),$$
$$W = W. * (XH^T)./(WHH^T),$$

where $.*$ and $./$ denotes the element-wise multiplication and division between a pair of matrices, respectively.

4. NMF for Topic Detection

4.1 Topic Detection via Factorization of User Participation Matrix

In the study of online forum, users' participation in a discussion can be reflected by their initiative and the follow-up posts to the discussion. Hereby we can obtain a nonnegative matrix containing the participation frequencies of each user in each discussion. A matrix X in size $n \times m$ thus means that n discussions are participated by m users at different frequencies.

Similar to the $tfidf$ normalization steps for document-word matrices, a $pfidf$ procedure is applied to the discussion-participation frequency matrices. This is motivated by the Zipf's law of user frequency shown in earlier section. The pf stands for participation frequency, and idf for inverse discussion frequency. If the entry $x_{ij} \in X$ denotes the raw participation frequency of user j in the discussion i, the $pfidf$ score is updated in the following manner,

$$x_{ij} = (pf_{ij}) \cdot (idf_j) = x_{ij} \cdot \log(\frac{n}{|d : d \ni user_j|}),$$

where $|d : d \ni user_j|$ means the number of discussions in which the user j participates. Furthermore, the discussion i is normalized to unit Euclidean length by dividing the L_2 norm of the $pfidf$ vector corresponding to the discussion i.

To detect the k groups of latent topics in discussions, the weighted discussion-participation matrix X can be factorized via NMF into a matrix W of $n \times k$ and a matrix H of $k \times m$. The two matrices after factorization have the effect of indicating the cluster membership. The cluster membership c_i of the $i-$th discussion is simply given by

$$c_i = \arg\max_j W_{ij},$$

where j is the label of the latent topic of the discussions. Usually the number of latent topics are a much smaller number than the total number of discussions. In our analysis, the number of latent topics ranges from three to fifteen while the total number of discussions are over one thousand.

4.2 Joint Factorization of Discussion-Participation and Discussion-Word

In topic detection or discussion clustering, in addition to the discussion-participation frequency matrix, a discussion-word matrix can be formed. For a set of n discussions where m words appear altogether, we can represent them into a matrix F of $n \times q$.

A simple objective to factorize both matrices can be formulated,

$$X = WH, F = WG,$$

where X and F are factorized to the same matrix W, together with H and G, respectively. And the objective function is

$$\min_{W,H,G\geq 0,} (||X - WH||^2 + \lambda||F - WG||^2), \qquad (1)$$

where λ is a user-specified constant. This leads to the following updating rules,

$$H = H. * (W^T X)./(W^T WH),$$

$$G = G. * (W^T F)./(W^T WG),$$

$$W = W. * (XH^T + \lambda FG^T)./(W(HH^T + \lambda GG^T)),$$

which can guarantee to keep the objective value non-increasing through the proof of trace operation and Lagrange transform. In addition to the updating rules, the following two separate updating rules are adopted as initialization steps,

$$W = W. * (XH^T)./(W(HH^T)),$$

$$W = \lambda W. * (FG^T)./(W(GG^T)).$$

5 Experiment

5.1 Data Extraction

This section describes the discussion clustering of a local popular web forum which provides a discussion cyberspace for people interested in Audio-visual equipments, in particular the higher-end or high fidelity equipments.In this forum, three distinct discussion boards are available to public users with assigned alias $AvBoard$, $ChatBoard$, and $2ndHandBoard$. In the first of the experiment, we conduct topic detection in $AvBoard$ using discussion participation data only.

In the second part of experiment, the discussions in the three boards are then merged together to form a testing dataset and the goal is to classify them back to their original boards. In the first setup, the attribute vector is formed, in bag-of-words sense, by only considering the words appearing in the subject field of a discussion. For the second attribute vector setup, each discussion is represented by a participation frequency vector. This vector has the number of components equals to the number of participators and each component corresponds to a specific participator. The value of a component is the number of messages that participator posts for that discussion. Both data are represented in matrix form and normalized in the way discussed in Section 4.1.

5.2 Topic Detection in Online Forum

In this section, we present results of topic detection using user participation only in the $AvBoard$. The pfidf-weighted user participation frequency matrix in Section 4.1 is decomposed using NMF into ten groups. The ten groups are evaluated by human expert as well as cluster entropy. Human expert evaluations of the latent topic nature of the clusters are shown in Table 1. Clusters that do not have coherent topics in the discussions are labeled as miscellaneous.

The latent topics discovered matches much close to the posting characteristics. Another important aspect of the results is that some of the latent topics discovered does not exist in the list of categories provided by the designer of the forum. For example, C1 of the Exhibition and shows about AV equipment does not exist in the forum. The same is also true for C2 and C3 which is related to do-it-yourself (DIY) in audio hobby. As DIY discussions on audio equipment is a hobby where much support and discussions are generated, the existence of sub-community based on it is quite evident. Furthermore, vintage equipment also found itself in a special sub-community. For more details of the clustering results, please visit http://www.comp.hkbu.edu.hk/˜chli/wi07data.html

Table 1. Latent Topics discovered in the AvBoard

C1	Exhibitions, shows
C2	Tube/DIY
C3	DAC DIY
C4	Compact disc
C5	Turntable, Vinyl discs
C6	Vintage Equipment
C7	Miscellaneous
C8	Japanese product
C9	CD players
C10	Miscellaneous

for clustered English titles translated from and http://www.comp.hkbu.edu.hk/˜chli/wi07cn.html

To measure how coherent are the discussions relative to the clusters, we measure the entropy of each cluster by normalizing the sum of each rows in W to be 1 and thus consider each row in W as proportional to the probabilities of the latent classes. Figure 3 shows the average entropies of discussions in the 10 clusters. Cluster 7 and 10 which are classified by expert to be miscellaneous have higher entropies than other clusters.

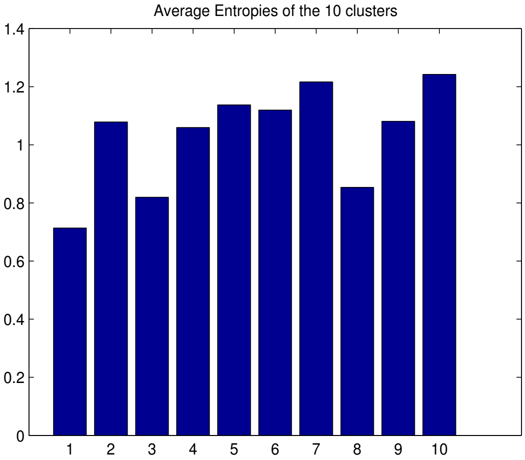

Figure 3. Distribution of Frequency of Post

Furthermore, as the complementary matrix H contains the coefficients of the user's contributions to the latent topics, the user's interest and role in the clusters can also be readily obtained. For example, in cluster one, the largest 10 user coefficients are plotted in Figure 4. There is a very clear leader in this cluster where the leader's coefficient are significantly larger than the second one. Furthermore, if we collect the overall statistics of post from this user, it is found that the post volume is only 0.4% and the overall post rank is 30. The same trend can also be found in several well-defined clusters. This agrees very well with previous social network results where leaders do not necessary post a lot but their influences are profound and can often arouse participation around them.

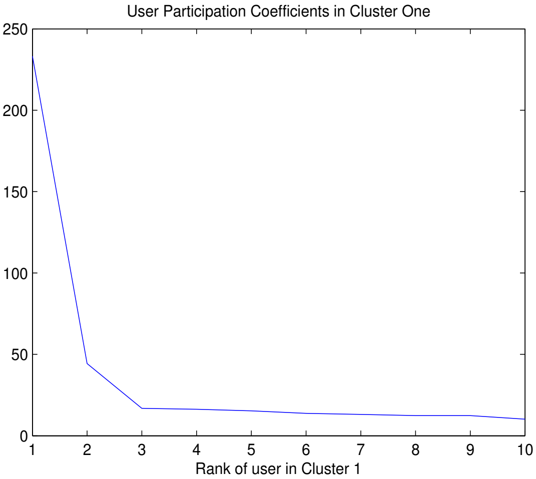

Figure 4. User Participation Coefficients in Cluster One

5.3 Clustering

For the clustering of the three boards of discussions, the data set contains 1003 discussions from $ChatBoard$, 1069 from $2ndHandBoard$, and 1040 from $AvBoard$. There are 7728 participators and 24791 words in total.

The clustering performance is measured by weighted purity. For a p-cluster task, if the factorization matrix W is $m \times k$ and hereby divides the dataset into k groups, the purity is calculated by first counting for each of the k groups the number of points with their true clustering label dominant in this group, and then divided by the total number of data point in the dataset. When $p = k$, this measure is equivalent to the typical clustering accuracy measure.

Table 2 shows the clustering purity on different k, while DPDW refers to the NMF utilizing both the discussion-participator and discussion-word matrices, and DP and DW refers to the NMF utilizing the discussion-participation and discussion-word matrix respectively. It can be noticed from the result that the clustering results can be significantly increased with the joint factorization approach.

Table 2. Purity Measure of AV Web Forum Clustering

k	DPDW	DP	DW
3	0.7382	0.5480	0.4968
6	0.8661	0.5646	0.6786
9	0.8746	0.5883	0.6735
12	0.8878	0.6071	0.6702
15	0.8668	0.6199	0.6591

6 Conclusion

User participation in online forum is essential to the analysis of online discussions and follows the Zipf's law. We presented methods for detecting topics based on discussion-participation, and also on both discussion-participation and discussion-word. Results of topic detection in online forum shows that the approach is feasible and latent topics previously unknown to the forum can be discovered. It should also be noted the degree of effectiveness could be dependent on the nature of the online forum. Furthermore, we also present results on integrating the use of document corpus with user participation to cluster discussions from several different discussion boards.

References

[1] J. Bishop. Increasing participation in online communities: A framework for human-computer interaction. *Comput. Hum. Behav.*, 23(4):1881–1893, 2007.

[2] P. J. Carrington, J. Scott, and S. Wasserman, editors. *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005.

[3] C. Ding, X. he, and H. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the Fifth SIAM International Conference on Data Mining*, pages 606–610, Newport Beach, 2005.

[4] P. Kollock. The economies of online cooperation: Gifts and public goods in cyberspace. In M. Smith and P. Kollock, editors, *Communities in Cyberspace*. Routledge, London, 1999.

[5] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[6] R. D. Nolker and L. Zhou. Social computing and weighting to identify member roles in online communities. In *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 87–93. IEEE Computer Society, 2005.

[7] U. T. Pentti Paatero. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[8] M. Richardson and P. Domingos. Markov logic networks. *Mach. Learn.*, 62(1-2):107–136, 2006.

[9] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003.

[10] G. K. Zipf. *The Psychobiology of Language*. Houghton-Mifflin, New York, 1935.