Social and Semantics Analysis Via Non-negative Matrix Factorization *

Zhi-li Wu
Computer Science Department
Hong Kong Baptist University

Chi-wa Cheng
Computer Science Department
Hong Kong Baptist University

Chun-hung Li
Computer Science Department
Hong Kong Baptist University

{vincent,victor,chli}@comp.hkbu.edu.hk

ABSTRACT

Social media such as Web forum often have dense interactions between user and content where network models are often appropriate for analysis. Joint non-negative matrix factorization model of participation and content data can be viewed as a bipartite graph model between users and media and is proposed for analysis social media. The factorizations allow simultaneous automatic discovery of leaders and sub-communities in the Web forum as well as the core latent topics in the forum. Results on topic detection of Web forums and cluster analysis show that social features are highly effective for forum analysis.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*; H.3.1 [**Information System**]: Content Analysis and Indexing; H.5.4 [**Information System**]: Information interfaces and presentation—*hypertext/hypermedia*

General Terms

Theory, Algorithms

Keywords

Social Network Analysis, latent topic detection, latent interest detection

1. SOCIAL INTERACTION MODEL

1.1 Latent Topic Detection via Factorization of User Interaction Matrix

In the study of social media involving dense interactions between users, relationships between the media and the individuals are often essential to the understanding of both the topics of discussion and the interest of users. Figure 1 shows the relationship between web media and individuals using the web site. In the case of Web forum, online discussion is the media where individuals participate according to their interest. By measuring the participation frequencies of each user in each discussion, and denoting it as matrix X of size $n \times m$, where n discussions are participated by m users.

*This work is partially supported by FRG of HKBU

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

This matrix can also be viewed as a bipartite graph where one set of nodes corresponds to the discussions and one set of nodes to the users.

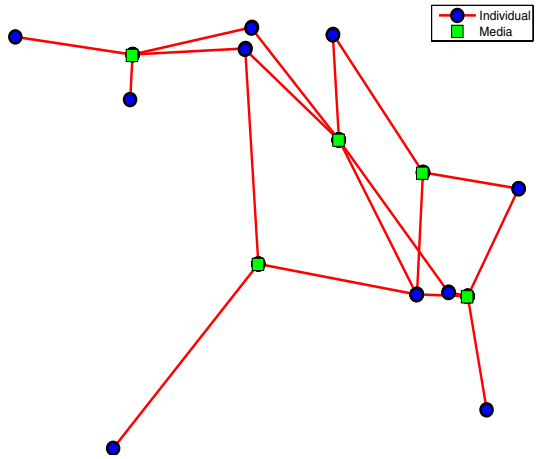

Figure 1: Social Interaction of Individual and Web Media

To detect the k groups of latent topics in discussions, the weighted discussion-participation matrix X can be factorized via non-negative matrix factorization (NMF) into a matrix W of $n \times k$ and a matrix H of $k \times m$. NNF has found lots of applications in text mining [3],[4]. 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.

1.2 Joint Factorization of Social and Semantic Attributes

For a set of n discussions where m words appear altogether, we can represent the discussions-words into a matrix F of $n \times q$. To factorize both matrices,

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

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

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

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,

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

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

2. EXPERIMENT

2.1 Data Extraction

A popular web forum in high fidelity Audio-visual equipments is studied. 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.

2.2 Topic Detection in Web Forum

The results of topic detection using user participation only in the $AvBoard$ is shown here. The pfidf-weighted user participation frequency matrix 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

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

match well to the posting characteristics. New latent topic, e.g. C1 of the Exhibition and shows about AV equipment that 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. The entropy measures of the identified cluster agrees very well with human evaluations where miscellaneous clusters have higher entropies.

2.3 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: 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

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.

3. CONCLUSION

User participation in Web forum is essential to the analysis of Web discussions. We presented methods for detecting topics based on discussion-participation, and also on both discussion-participation and discussion-word. Results of topic detection in Web 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 Web forum. Furthermore, we also present results on integrating the use of document corpus with user participation to cluster discussions from several different discussion boards.

4. REFERENCES

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

[2] 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.

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

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