Wavelet-Based Clustering of Social-Network Users Using Temporal and Activity Profiles

Lipika Dey and Bhakti Gaonkar

Innovation Labs, Tata Consultancy Services, Delhi, India
(lipika.dey, bhakti.g)@tcs.com

Abstract. Encouraged by the success of social networking platforms, more and more enterprises are exploring the use of crowd-sourcing as a method for intra-organization knowledge management. There is not much information about their effectiveness though. While there has been some emphasis on studying friend networks, not much emphasis has been given towards understanding other kinds of user behavior like regularity of access or activity. In this paper we present a wavelet-based clustering method to cluster social-network users into different groups based on their temporal behavior and activity profiles. Cluster characterization reveals the underlying user-group characteristics. User data from web and enterprise social-network platforms have been analyzed.

Keywords: Social Network Analysis, Wavelet transformation, Hierarchical K-means clustering.

1 Introduction

Social networking sites like Twitter and Facebook are figuring in an increasingly important way to market researchers due to their astronomically rising user base over a very short time. Efforts are on to tap the content that is generated in these sites to get valuable business insights. There is also an attempt to replicate these platforms within the secure environments of enterprises and motivate employees to engage in similar social interactions. It is expected that these platforms can motivate knowledge and expertise sharing in a more pro-active and effective way. However, these initiatives can only succeed if user participation is considerably steady and continuous, which is not necessarily true. Studies show that though popular social network sites like Twitter have an increasing user base, but there is also a huge attrition rate. Characterization of users based on their temporal behavior and activity profiles can not only help in identifying valuable and active user groups, but also help in improving the effectiveness of such platforms by identifying rare or infrequent users and motivating them to participate more through various incentives. Identifying anomalous behavior can be yet another important application of this.

In this paper, we present an efficient way of clustering users into different groups based on their temporal behavior in social network forums. While there has been a lot of focus on studying user-groups based on content generated by them, not much work exist on analyzing their temporal behavior. We show that wavelet-based analysis can

effectively identify different categories of users, whose behaviors can be easily characterized. The rest of the paper is organized as follows. Section 2 surveys related works. Section 3 presents the background of wavelet analysis. The details of our wavelet based clustering algorithm are discussed in Section 4. Section 5 describes the experiments conducted and the results obtained.

2 Survey of Related Work

Wavelet transform being spatial-temporal in nature can capture temporal as well as frequency-based behavior of signals. In [1], a technique based on Discrete Wavelet Transform was proposed to cluster ECG signals. Though wavelet based analytical techniques are very powerful, they have not seen much use in the field of social-network analysis. There have been several earlier attempts to cluster web-site users into different groups based on their access patterns. In [2] a novel approach was presented to cluster user profiles based on mass distribution using Dempster-Shafer's theory. Profile similarity was used to make recommendations, personalize Web sites, and create targeted advertisements. In [3], a trace analysis was conducted to characterize user access to videos over web (VOW). Their analysis revealed interesting discoveries like requests for same video spanned over short span of time and that small number of machines accounted for large number of overall requests. In [4], a very recent work, it has been shown how content popularity grows and fades over time in Twitter. In order to uncover the temporal dynamics of online content this has been formulated as time series clustering problem using a similarity metric that is invariant to scaling and shifting. K-Spectral Centroid (K-SC) clustering algorithm is proposed, that can effectively find cluster centroids with their similarity measure applied on wavelet coefficients of the time series.

3 Brief Overview of Wavelet Analysis and Wavelet Transformation

Wavelets are mathematical functions that are very useful in representing data due to various properties like compact support, vanishing moments, dilating relations etc. [5, 6, 7]. Compact support guarantees the localization of wavelets, while vanishing moment guarantees that wavelet processing can distinguish the essential information from non-essential information. These features including facilities for hierarchical representation and manipulation make wavelets a very powerful tool. Wavelet transforms are spatial-temporal in nature. They capture the time as well as the frequency of the signals. In wavelet analysis, the scale that is used to look at data plays a special role.

The Discrete Wavelet Transform can be described as a series of filtering and sub sampling. In each level in this series, a set of 2^j-1 coefficients are calculated, where $j < J$ is the *scale* and $N = 2^J$ is the number of samples in the input signal. The coefficients are calculated by applying a high-pass *wavelet filter* to the signal and down-sampling the result by a factor of 2. At the same level, a low-pass *scale filtering* is also performed (followed by down-sampling) to produce the signal for the next

level. Each set of scale-coefficient corresponds to a “smoothing” of the signal and the removal of details, whereas the wavelet-coefficients correspond to the “differences” between the scales.

Given a signal f_0 having dimension D_f , low frequency filter L and a high frequency filter H , the Fast wavelet transformation of the signal is computed as follows:

1. Apply H to f , returning a new signal f_H , with a domain half the size of D_f ,
2. Similarly, apply L to f , returning another new signal f_L , with a domain half the size of D_f ,
3. If size of D_f is 2, then return $[f_H, f_L]$
4. Apply Wavelet transform (recursively) on f_L ,
5. Return f_H concatenated with the result.

The resultant transformed signal f_0 is $(Lf_k, Hf_{k-1}, Hf_{k-2}, \dots, Hf_0)$ where Lf_k is the wavelet coefficient of signal f_0 at the k^{th} (coarsest) level. It contains the approximated information about original signal. Hf_j are the wavelet coefficients of the signal at j^{th} level containing the detailed information of the signal at that level.

The simplest wavelet transformation function is Haar transform which has a single vanishing moment. Haar transform performs an average and difference on a pair of value. We have used Daubechies2 wavelet transforms having two vanishing moments. The corresponding low-level and high-level filters are given by (0.12940952, 0.22414386, -0.8365163, 0.4829629) and (0.4829629, 0.8365163, 0.22414386, -0.12940952) respectively.

4 Wavelet-Based Clustering of User Interaction Data

We assume that user interaction data is represented as a time-series, which records the activity of the user at regular time-intervals. Measure of activity in a given time interval is the number of times the user repeats the activity in that period. Activity definition depends on the type of social network and may include actions like asking a question or answering one, sending messages or tweets, or simply updating status.

Grouping users based on their behavior is performed in two phases. In phase I, Fast Daubechies Transform is applied to the time series data. This transformation yields a set of values that are linear combinations of the original elements of the time series. As mentioned earlier, the coarsest level coefficients obtained with the transformation can provide approximate information about the users, while at each higher level more detailed information about the behavior can be captured. In phase II, we exploit this property of wavelet coefficients to refine the process of clustering repeatedly and obtain more refined and homogeneous groups.

The hierarchical clustering algorithm has been designed based on *k-means*. Initially, the coefficients obtained at the coarsest level are segregated into k clusters. Phases II is repeated for each new cluster, till the most detailed coefficients are taken into account. The clustering process is summarized below

Wavelet based Hierarchical Clustering Algorithm

Let n denote the number of users and m denote the number of time-intervals considered

Input: Set of signals $\bar{U}_n$, where $u_n = \langle f_i(t_j) \rangle$, $i = 1, 2 \dots n$ and $j = 1, 2 \dots m$. Since m should be a power of 2, let $m = 2^s$. Extra fields in the signal, if needed, are padded with zeroes.

- Parameters $k_1, k_2, \dots, k_s$, where k_t is the number of clusters to be created during the t^{th} iteration.
 - Output: p clusters, where $p = k_1 * k_2 * \dots * k_m$
1. Apply Fast Daubechies Transform (FDT) on each element u_n of $\bar{U}_n$. The data set is transformed to $\bar{W}_n$, where $W_i(t_j) = , W(f_i(t_j)) = W_i(t_j)$
 2. Assign $C_{s+1}^1 = W_i(t_j)$
 3. For $t = s, s-1, \dots$,

For $c=1, 2, \dots, k_1 * k_2 * \dots * k_{s-1-t}$
 - a. Apply K-means clustering on set of elements belonging to C_{t+1}^c considering only t^{th} level coefficients of the elements w_n belonging to C_{t+1}^c , dividing each cluster into k_t clusters.
 - b. As a result of step a. $k_1 * k_2 * \dots * k_t$ are formed.

5 Experiments and Results

Several experiments were conducted to study users of different types of social networks. The first and second dataset comprise of Twitter users for a period of one month in 2009 and fifteen days in 2010 respectively. The third and fourth datasets comprise data from an Enterprise Social Network Portal (ESNP) where users ask questions or answer questions posted by other users. It consists of employees and their monthly behavior over 25 months from April 2008 to April 2010. The Twitter data time-series contain the number of tweets sent by each user on the different days. The Enterprise data contains number of questions or answers posted by users in a given month. Table 1 presents details of the datasets as well as the results obtained.

It can be observed from Table 1 that the distributions are identical with one large cluster receiving majority of the data points. It is found that the major cluster always contains infrequent users with low activity profiles. This cluster emerges at the end of first level itself, and remains unchanged through the next iterations. Figure 1 shows the clusters formed for the different datasets after the first level of clustering, i.e. at the coarsest level of approximation. These plots have been generated using total activity during the period. The last column of Table 1 shows that the percentage of users exhibiting significant volumes of activity and continuous presence is very low.

Table 1. Details of Datasets and Clustering results

Dataset	Number of users	Time Period	Clusters (#)	% users in largest cluster	% users with high-activity
Twitter I	200000	31 days	48	54	0.16
Twitter II	44895	16 days	16	63	0.27
ESNP Answers	21420	25 months	32	93	0.47
ESNP Question	19810	25 months	32	78	0.35

Fig. 1. [a] Twitter data I [b] Twitter data II [c] ESNP Answers [d] ESNP Questions

Figures 2[a] and 2[b] show activity profiles of sample users from the major cluster (red dots) and one minor cluster (blue dots) from second Twitter and ESNP answers data sets respectively. In figure 2[a], red dots depict low-profile users with an average activity of 2 tweets and an average presence of 1.6 days over the period of 16 days. Blue dots in the figure depict highly active users with an average activity of 4.5 tweets and an average presence of 12 days. Similarly in figure 2[b], red dots depict low-profile users with an average activity of 2.7 answers and an average presence of 3 months, over the period of twenty five months. Blue dots in the figure present high-profile users with an average activity of 30.7 answers and an average presence of 11.2 months. Figure 2c analyses the content of ESNP high activity users and will be discussed later.

Fig. 2. Inactive and most active users [a] Second Twitter dataset [b] ESNP answers [c] Category distribution of content posted by most active ESNP users

It is obvious that majority of social network users are not regular and maintain a low profile. Analysis of Twitter user profiles further reveal that users with highest

levels of activity are almost always news groups. In both the datasets, the most active group consisted of IDs of global news agencies like *breaking news* and *TOInews*, while the second most active set of users included regional news agencies like *AtlantaNews*, *TOIMumbai*, *TOIDelhi* etc. It is interesting to observe that the patterns of users and usage have remained same over the years, in spite of growing number of users. Further, analysis of ESNP data showed that around 57% of users who logged into the network on any given month, did not do so the very next month. Since very few users take an active role on a continuous basis, the purpose of such networks may not be fulfilled.

The ESNP data used for this experiment had user-given category tags Technology, Programming Languages, HR, Finance, Fun and Entertainment etc. associated to the content. Figure 2c shows that bulk of content contributed by the most active users belong to the Fun and Entertainment category. The other popular categories are culture and Work Life balance. Thus it can be concluded that social networks are positively contributing towards employee bonding. But more can be done to make social networks a platform for enterprise knowledge sharing effectively.

6 Conclusions

In this paper we have explored the temporal behavior of users in social network. We have proposed an effective way of clustering users into different groups and characterized them. It is found that majority of users in any social network are low-profile and irregular. Within enterprise, active interaction is mostly restricted to non-technical content. Schemes to engage users effectively are also essential. Introduction of proper incentives are needed to make social network a viable platform for knowledge sharing. Further work is being carried on to analyze the content and its effectiveness for enterprise.

References

1. Ramakrishnan, A.G., Saha, S.: ECG coding by wavelet-based linear prediction. *IEEE Trans. on Biomedical Engineering* 44, 1253–1261 (1997)
2. Acharya, S., Smith, B., Parnes, P.: Characterizing User Access to videos on the World Wide Web. In: *Proceedings of MMCN* (2000)
3. Xie, Y., Phoha, V.V.: Web user clustering from access log using belief function. In: *Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001)*, pp. 202–208. ACM Press, New York (2001)
4. Yang, J., Leskovec, J.: *Patterns of Temporal variation in Online Media*
5. Graps, A.: *An introduction to Wavelets*. IEEE, Los Alamitos (1995)
6. Li, T., Li, Q., Zhu, S., Ogihara, M.: A Survey on wavelet Application in Data Mining. *ACM SIGKDD Exploration Newsletter* 4(2) (2002)
7. Daubechies, I.: *Ten lectures on Wavelets*. SIAM, Philadelphia (1991)