EXPLORING TEMPORAL ASPECTS IN USER-TAG CO-CLUSTERING

E. Giannakidou[*], *V. Koutsonikola, A. Vakali*

Aristotle University of Thessaloniki
Department of Informatics, Greece

I. Kompatsiaris

Centre of Research & Technology Hellas
Informatics and Telematics Institute, Greece

ABSTRACT

Tagging environments have become an interesting topic of research lately, focused mainly on clustering approaches, in order to extract emergent patterns that are derived from tag similarity and involve tag relations or user interconnections. Apart from tag similarity, an interesting parameter to be analyzed during the clustering/mining process in such data is the actual time that each tagging activity occurred. Indeed, holding a temporal dimension unfolds macroscopic and microscopic views of tagging, highlights links between objects for specific time periods and, in general, lets us observe how the users' tagging activity changes over time. In this article, we propose a time-aware user/tag clustering approach, which groups together similar users and tags that are very "active" during the same time periods. Emphasis is given on using varying time scales, so that we distinguish between clusters that are robust at many time scales and clusters that are somehow occasional, i.e. they emerge, only at a specific time period.

1. INTRODUCTION

With the advent of Web 2.0, tagging practices constitute a collective fashion of metadata creation, which, especially in the case of digital content retrieval, is very important. As more and more people have supported this surge, tagging provides a rich knowledge source to study social patterns and emergent drifts/ directions in the web user community. Further, the fact that tags are applied on a daily basis gives this data source an extremely dynamic nature that reflects the changes and the evolution of community focus. Therefore, a temporal analysis of tag data may provide insight regarding a topic or trend evolution over time at a macroscopic level or a microscopic level.

While there has been substantial research on overcoming tags' limitations (such as ambiguities, synonym words, etc) and analyzing tag patterns to infer information about the user community, most of the current approaches rely on a static basis ([1, 2]) and there is a limited number of efforts that study the evolution and dynamics in tagging activities, by using an explicit temporal dimension. Existing approaches, in this area, employ statistical methods on tags' time distributions, to identify emerging trends or events of interest ([3, 4]). The same kind of analysis may be also utilized for extracting tags' dynamics models and apply them on tag suggestion techniques [5]. The idea of analyzing temporal tagging patterns to induce time-aware user profiles was first introduced in [6]. However the analysis in this approach was based on predefined timeframes.

In this paper, we use time-aware co-clustering to analyze tagging data and obtain groups of like-minded users as expressed through their tagging activity. Clustering has been widely used in social tagging systems to support several applications, like information retrieval, providing recommendations, or the establishment of user profiles and the discovery of topics. However unlike other approaches, here we stress the importance of temporal information embedment in tagging data analysis. The work was inspired by the fact that, in practice, users exhibit varying tagging behavior at different timescales (e.g. on a yearly, monthly or daily basis). Moreover, clustering digital objects by time can be applied successfully for event detection [7]. Therefore, the consideration of time, along with tags preferences is substantial in clustering, since, in fact, the time in which users perform certain tagging activities is a crucial criterion for characterizing their particular needs and preferences. The rest of the paper is organized as follows. In the next section a detailed analysis of the proposed approach is given, including similarity measures and pseudocode-description of the proposed algorithm. Evaluation using a Flickr dataset is presented in Section 3. In Section 4 some conclusions and future work are discussed.

2. CAPTURING USER/TAG TIME-AWARE PATTERNS

2.1. Problem Formulation

The fact that most times the tagging activity of a user expresses the user's personal viewpoint and interests reveals an interconnection between users and tags. This poses a duality between user and tag clustering. Such a problem was discussed in [8], where the idea of co-clustering simultaneously

[*] E. Giannakidou is also with the Informatics and Telematics Institute, Greece.

items of different datasets was proposed. To this end, here, we use a co-clustering method that yields a series of clusters, each of which contains a set of users together with a set of tags. The cluster assignment is defined by a $Similarity$ function, that analyzes tagging activity patterns and searches for commonalities. More specifically, co-clustering users and tags based on tagging activity patterns is faced as a twofold problem that: i) deals with the tags preferences, and ii) identifies the temporal patterns involved in the tag usage assignment. Thus, the proposed algorithm should address the above two criteria in the $Similarity$ function.

The first track of the problem involves the similarity between a user and a tag in a conceptual framework. Thus, for a tag and a user to be considered as highly similar, it is a prerequisite that the user has used this particular tag or another one semantically close to it. To calculate this kind of similarity between a user u_x and a tag t_y (called, hereafter, as *Semantic Similarity*, *SeS*), we use the WordNet lexicon, as follows:

$$SeS(u_x, t_y) = \max_{t_z} \frac{2 \times depth(LCS)}{[depth(\overrightarrow{t_z}) + depth(\overrightarrow{t_y})]}, \qquad (1)$$

$\forall t_z$ assigned by u_x, where $depth(\overrightarrow{t_x})$ is the maximum path length from the root to $\overrightarrow{t_x}$ and *LCS* is the least common subsumer of $\overrightarrow{t_x}$ and $\overrightarrow{t_y}$[9].

The second track of the problem examines the time locality between a user's tagging behavior and a tag's usage patterns. The goal here is to bring together users and tags that have similar usage patterns over time. To this end, we divide the entire time period into I sequential timeframes of size τ and examine the patterns of each user and each tag in each of the underlying timeframes. In our analysis, we use the vector-space model for a user's u_x and a tag's t_y temporal representation, as follows:

$$u_x = [u_{x1}, u_{x2}, \ldots, u_{xI}], x = 1, \ldots, U,$$

$$t_y = [t_{y1}, t_{y2}, \ldots, t_{yI}], y = 1, \ldots, T,$$

where u_{xj} and t_{yj} are the number of tags user u_x has assigned and the number of times the tag t_y has been used, respectively, during the timeframe j, $j = 1, \ldots, I$. By calculating the inner product between a u_x vector and a t_y vector, we obtain the similarity between the two vectors, which corresponds to the temporal locality of the specified user and tag (called, hereafter, as *Temporal Similarity*, *TeS*). Thus,

$$TeS(u_x, t_y) = \frac{\sum_{k=1}^{I} u_{ik} \cdot t_{jk}}{\sqrt{\sum_{k=1}^{I} u_{ik}^2 \cdot \sum_{k=1}^{I} t_{jk}^2}}, \qquad (2)$$

Emphasis is given on capturing time locality at varying time scales for tracking clusters that are occasional (i.e. exist only at specific time-scale analysis) and clusters that are regular and robust at many time scales. To achieve this, we experiment with various values of τ.

The total similarity between a user u_x and a tag t_y is estimated by considering both their semantic and temporal similarities (Equations 1,2) as follows:

$$Similarity(u_x, t_y) = SeS(u_x, t_y) \cdot TeS(u_x, t_y)$$

The values of $Similarity$ function between each of the u users and t tags are then used ro form $u \times t$ table **Sim** as follows:

$$\mathbf{Sim}(x, y) = Similarity(u_x, t_y),$$

where $x = 1, \cdots, u$ and $y = 1, \cdots, t$.

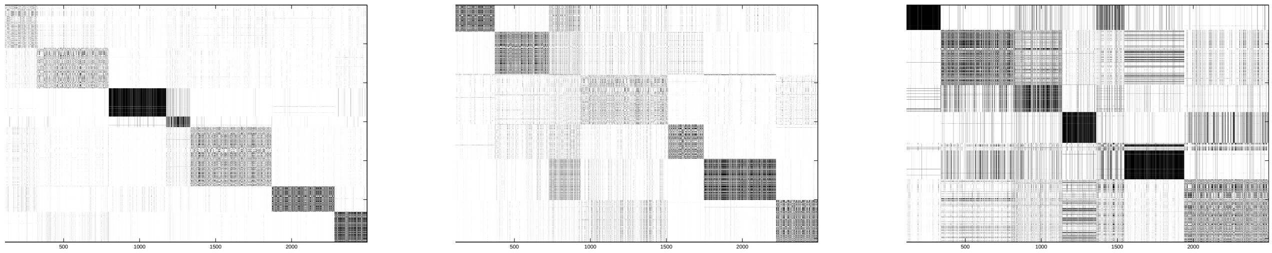
2.2. Time-aware Co-clustering Algorithm

Co-clustering extends traditional spectral clustering algorithms for grouping together elements from different data sources ([8]). This idea was applied in a web 2.0 environment, in [10] for obtaining joint groups of tags and resources, to improve retrieval of resources by exploiting their relation to tags. Another approach for using spectral clustering, still in web 2.0 context, is presented in [11], for capturing the three dimensions in social tagging data (i.e. tags, users and resources) and combining multiple values of similarity to get groups of related items.

Algorithm 1 The CO-CLUSTERING algorithm.

```
Input: The set $U$ of $u$ users and the set $T$ of $t$ tags over a
    time period $\mathfrak{T}$ and an integer $k$
Output: Multiple sets $C = \{C_1, \ldots, C_k\}$ of $k$ subsets consisting of elements from both $U$ and $T$, that have the same
    patterns at the underlying timeframe duration $\tau$.
/*Preprocessing*/
$U^* = Preprocess(U)$
$T^* = Preprocess(T)$
/*Capturing similarities at different time-scales*/
for $\tau$ in $[1, \mathfrak{T}]$ do
    $TeS = CalculateTemporalSimilarity(U^*, T^*, \tau)$
    $SeS = CalculateSemanticSimilarity(U^*, T^*)$
    $Sim = TeS \bullet SeS$
    /*Co-clustering process*/
    $(D_u, D_t) = ComputeDegreeTables(Sim)$
    $NSim = D_u^{-1/2} Sim D_t^{-1/2}$
    $(L_u, R_t) = SVD(NSim)$
    $SV = CreateIntegratedTable(D_u, D_t, L_u, R_t)$
    $C = k - means(SV, k)$
end for
```

Here, we employ the method presented in [10] and use the similarity matrix, **Sim** to yield time-aware $k-$partitionings of users and tags. As it has been proven in [8], the k left and right singular vectors of an appropriately scaled similarity matrix $NSim = D_u^{-1/2} SimD_t^{-1/2}$ provide a real approximation to the $k-$partitioning problem. The D_u and D_t are the diagonal degree tables of users and tags respectively. Let L_u denote



(a) time-aware user/tag co-clustering using semantic and temporal dimensions

(b) static user/tag co-clustering using only semantic dimension

(c) user/tag co-clustering using only temporal dimension

Fig. 1. Clusters' semantic and temporal coherence evaluation (darker blocks indicate more coherent clusters.)

the u x k table of the left singular vectors and R_t the t x k table of the right singular vectors of **Sim** table. In order to perform a simultaneous clustering of u_i, $i = 1, \ldots, u$, and t_j, $j = 1, \ldots, t$, elements, we create the $(u + t)$ x k two dimensional table *SV* defined as:

$$SV = \left[\begin{array}{c} D_u^{-1/2} L_u \\ D_t^{-1/2} R_t \end{array} \right]$$

Running a typical clustering algorithm on SV results in k clusters containing elements from both users and tags sets. A pseudocode description of the approach is presented above.

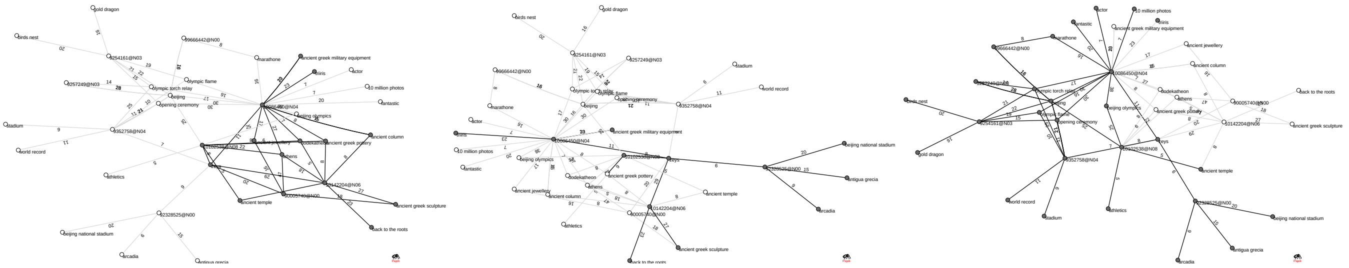
In the first step of the CO-CLUSTERING algorithm, a data preprocessing (line 2) takes place where a filtering of the tags is applied. More specifically, many users have the tagging habit to merge many tags into one single word, resulting, thus, into numerous meaningless compound terms. To tackle this, we analyze such terms and decompose them into their constituent elements-tags. After this metadata decomposition, we remove rare elements (users or tags), since typically such objects are considered trivial in tagging data analysis. The preprocessing step results into U^* and T^* sets of users and tags, respectively. Then, an iterative process occurs (lines 5-15), at each step of which we obtain a $k-$partitioning of U^* and T^* for a specific value of τ, i.e. at a particular time-scale (e.g. on daily, monthly scale, etc), based on the analysis that was described earlier. Each iteration finalizes with the k obtained clusters which contain both users and tags that have similar usage patterns over time at the current timescale τ (line14).

3. EXPERIMENTATION

We tested our method on a Flickr dataset of 1218 users, 6764 photos, and 2496 unique tags that span in a time period from Sep. 2007 to Sep. 2008. To demonstrate the compactness of the clusters in terms of semantic and temporal cohesion, we visualize the clusters by reordering the similarity matrix, so that same cluster entities are contiguous (in rows and in columns). The darker the coloring of a cell $c(i, j)$ where $1 \leqslant i, j \leqslant U + t$ the more similar the objects at position (i, j) are. Thus, clusters appear as symmetrical dark squares across the main diagonal. We conducted experiments gravitating semantic or temporal similarity or considering both of them equally, for various values of k and τ. In Figure 1 we indicatively present the clustering outline for $k = 7$ and $\tau = 10$, in three different cases. Particularly, the plot shown in (a) depicts the reordered similarity matrix **Sim**, after applying the proposed time-aware co-clustering, the plot shown in (b) depicts the reordered similarity matrix **TeS**, after applying a semantic based clustering, and the plot shown in (c) depicts the reordered similarity matrix **SeS**, after applying a time based clustering. It can be seen that the proposed method succeeds in finding dark rectangles across the diagonal, which indicates that the proposed similarity function guides the clustering process to the identification of coherent clusters in terms of both temporal and semantic dimensions (a). The compactness deteriorates a lot, in case we omit one dimension, as shown in (b), in which we omit the temporal dimension and visualize the temporal locality in clusters obtained by semantic-based clustering, and in (c), that depicts a semantic similarity visualization in time-based clustering i.e. the semantic dimension is not considered in the clustering.

Next, we want to show the impact the timeframe's size has on the clustering and we experimented with various values of τ. It is a fact that users' tagging behavior changes over time for a number of reasons, such as changing of interests, commenting on specific events, following trends, etc. Moreover, the tags' popularity may increase in specific timeframes, during which the tag expresses a current trend, and decrease in other timeframes, during which the trend starts to vanish. Therefore, changing the values of τ results in clusters of different memberships. More specifically, the outcome of our analysis is that by choosing small values of τ (e.g. 1, 10, 30, that is daily up to monthly timeframes), we extract occasional user interests. This kind of analysis may be exploited for event tracking or detection of ephemeral trends. On the other hand, larger values for τ unfold users' regular interests and matters that are of concern to the user community on a long-term basis. Table 1 depicts user/tag co-clustering assignments for 2 indicative users, for varying values of τ.

Further, integrating the temporal dimension in the mining process allows us to catch the users' traces in the web, accord-



(a) Olympics cluster in June 2008 (b) Olympics cluster in July 2008 (c) Olympics cluster in August 2008

Fig. 2. Olympics 2008 clusters at different time periods.

Table 1. Clusters' Assignment for varying values of τ.

User	$\tau = 30$	$\tau = 100$
User1	olympics2008, beijing, flame, opening ceremony	ancientgreece, acropolis, parthenon, archaeology, ancient-civilizations
User2	earthquake, china, disaster, ruin, disasterassistanceresponseteam	wedding organizing, party

ing to their tagging activity, and, at the same time, visualize the groups' evolving and changing. To do so, we represent a tagging environment as a network, where the nodes correspond to users or tags and the links denote that a user has assigned a tag. Figure 2 shows a cluster obtained for $\tau = 30$ on 3 different timeframes, June 2008 (a), July 2008 (b), August 2008 (c). The active/inactive nodes and links denote the presence/absence of activity in each specific timeframe. The cluster includes users and tags related to topics such as *ancient greece* and *olympics*. The massive tagging activity on Aug 2008 about *Olympics* is due to the Olympics 2008 event that attracted Olympics-friends to comment on it, through tags.

4. CONCLUSIONS

Changing patterns in networks over time show how networks form, grow and wane. By understanding such patterns in tagging networks, we can derive the potential causes and consequences of change and predict network evolution. In this paper, a time-aware co-clustering approach was presented that allows the extraction of such patterns, along with the ability to distinguish between users' regular and occasional interests. A more automatic analysis for defining the timeframe size, τ, is part of our future work.

5. ACKNOWLEDGMENTS

The work presented in this paper was partially supported by the European Commission under contracts FP7-215453 WeKnowIt and FP7-248984 GLOCAL: Event-based Retrieval of Networked Media.

6. REFERENCES

[1] P. Mika, "Ontologies are us: A unified model of social networks and semantics," in *4th ISWC*, Ireland, 2005, pp. 522–536.

[2] L. Specia and E. Motta, "Integrating folksonomies with the semantic web," in *4th ESWC)*, Austria, 2007, pp. 624–639.

[3] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme, "Trend detection in folksonomies," in *1st Int. Conf. on Semantics And Digital Media Technology*, 2006, vol. 4306, pp. 56–70.

[4] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *30th Int. SIGIR*, 2007, pp. 103–110.

[5] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *16th Int. Conf. on World Wide Web*, 2007, pp. 211–220.

[6] V. Koutsonikola, A. Vakali, E. Giannakidou, and I. Kompatsiaris, "Clustering of social tagging system users: A topic and time based approach," in *10th Int. Conf. WISE*, 2009, vol. 5802, pp. 75–86.

[7] M. Cooper et. al., "Temporal event clustering for digital photo collections," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2005.

[8] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *7th SIGKDD*, California, 2001, pp. 269–274.

[9] C. Fellbaum, *WordNet, an electronic lexical database*, The MIT Press, 1998.

[10] E. Giannakidou, V. Koutsonikola, A. Vakali, and I. Kompatsiaris, "Co-clustering tags and social data sources," in *9th Int. Conf. on Web-Age Information Management, China*, 2008, pp. 317–324.

[11] A. Nanopoulos, H. Gabriel, and M. Spiliopoulou, "Spectral clustering in social-tagging systems," in *10th Int. Conf. on Web Information Systems Engineering*, 2009, pp. 87–100.