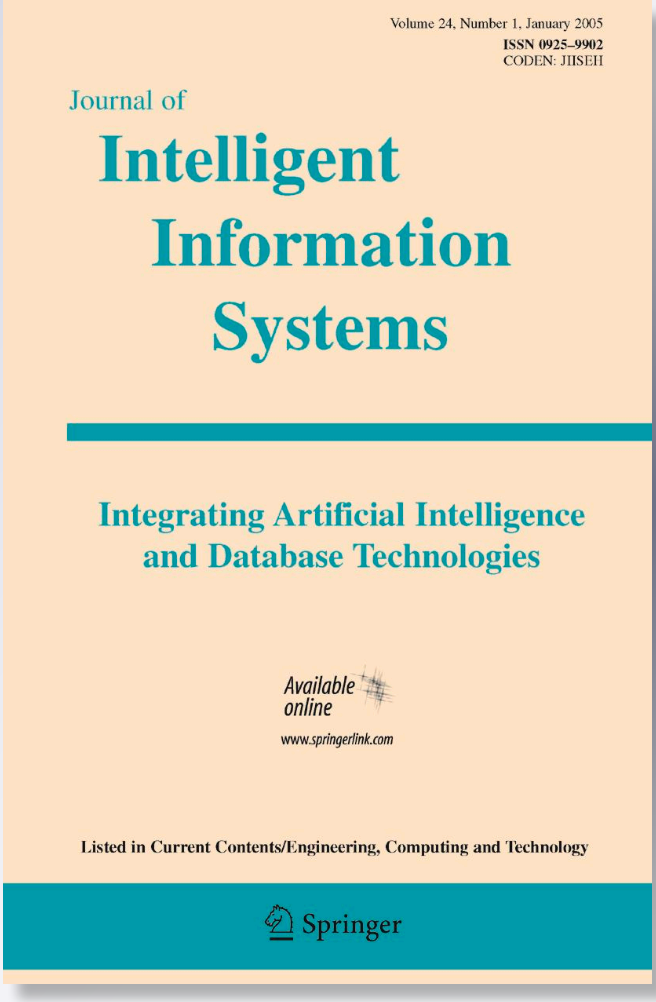
In & out zooming on time-aware user/tag clusters

Eirini Giannakidou, Vassiliki Koutsonikola, Athena Vakali & Ioannis Kompatsiaris

Journal of Intelligent Information Systems
Integrating Artificial Intelligence and Database Technologies

ISSN 0925-9902

J Intell Inf Syst
DOI 10.1007/s10844-011-0173-4



J Intell Inf Syst
DOI 10.1007/s10844-011-0173-4

In & out zooming on time-aware user/tag clusters

**Eirini Giannakidou · Vassiliki Koutsonikola ·
Athena Vakali · Ioannis Kompatsiaris**

Received: 18 January 2011 / Revised: 25 July 2011 / Accepted: 26 July 2011

Abstract The common ground behind most approaches that analyze social tagging systems is addressing the information challenge that emerges from the massive activity of millions of users who interact and share resources and/or metadata online. However, lack of any time-related data in the analysis process implicitly denies much of the dynamic nature of social tagging activity. In this paper we claim that holding a temporal dimension, allows for tracking macroscopic and microscopic users' interests, detecting emerging trends and recognizing events. To this end, we propose a time-aware co-clustering approach for acquiring semantic and temporal patterns out of the tagging activity. The resulted clusters contain both users and tags of similar patterns over time, and reveal non-obvious or "hidden" relations among users and topics of their common interest. Zoom in & out views serve as visualization methods on different aspects of the clusters' structure, in order to evaluate the efficiency of the approach.

Keywords Time-aware clustering · Social tagging systems ·
Users' interests over time · Events

1 Introduction

As more and more people adopt tagging practices, Social Tagging Systems (STSs) have become rich user-established repositories that enable the extraction of activity patterns reflecting current trends and interests within the users' community. A

E. Giannakidou (✉) · V. Koutsonikola · A. Vakali
Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece
e-mail: eirgiann@csd.auth.gr

I. Kompatsiaris
Informatics and Telematics Institute, CERTH, Thermi-Thessaloniki, Greece

Published online: 24 August 2011

number of techniques have been proposed in the literature for the processing and manipulation of these aggregated user input, most of which aim to extract such patterns, via clustering (Begelman et al. 2006; Specia and Motta 2007; Nanopoulos et al. 2009; Giannakidou et al. 2008). The majority of existing clustering approaches in STSs are based on tag similarity and some of the typically used metrics incorporate tag co-occurrence statistics and/or external knowledge resources (e.g. lexicons; Giannakidou et al. 2008). Indeed, since tags reflect the users' preferences and content semantics, patterns of similar elements in an STS can be derived from tag similarities.[1] The information derived from such approaches can be used for a number of applications, such as tracking users' needs & usage trends, tag patterning, profile learning, emergence of colloquial vocabularies, to name but a few.

The main drawback, though, of approaches based solely on tag similarity is that they are ineffective in capturing similarities based on users' activity over specific time frames (i.e. temporal segments), as they do not incorporate any time-related information. In most cases, a static view of the social tagging process is followed, aggregating the span of user activities in a single time span. In practice though, a user's tagging behavior, usually, changes over time. This may be due to a number of reasons, as for instance changing of interests, following trends, commenting on specific events and so on. It is a fact that the largest proportion of the data available in an STS is associated with an explicit temporal context, since each tagging task takes place at a particular time. Currently, such tagging timestamps are largely unexploited, while only a limited number of applications use time as an extra dimension in their analysis (see Section 4). Therefore, studying the tagging activity in a vast, evolving STS requires a different perspective, able to capture patterns that reflect changes in users' tagging activity over time. Such an approach is rather important in time critical applications, like trend recognition, advertising and recommendations which can certainly benefit from such a data analysis.

In this article, we claim that integrating temporal information in the mining/clustering process, along with a typically used tag similarity metric, is expected to yield more accurate and detailed view of the activity in an STS. Moreover, the inherent emphasis of such systems on users and the fact that the temporal dimension typically underlines the users' drifting towards certain topics over time constitute users, along with tags, to be two essential components to analyze simultaneously. For instance, a trend or an event of interest are imprinted by both users' and tags' dense activity patterns around a certain time period in an STS. In order to capture such knowledge, it is important to reveal the, often, hidden class-based correlation between users and tags. To this end, a time-aware co-clustering approach is proposed that allows to explore similar patterns of users and tags at different time scales. We argue that by understanding such patterns, we can unfold macroscopic (i.e. long-term) and microscopic (i.e. short-term) views of tag usage and users interests. It is like we use a time-sliding tool that scans the timeline of the tagging activity, in an effort to result with clusters that expose user choices at particular time frames. The extracted clusters consist of two types of objects, i.e. users and tags, and are expected to reveal users that share the same interests at the same time frames, as expressed

[1] As tag similarity refers mostly to tags' conceptual similarity, we will pertain to tag-based similarity as semantic similarity, throughout the article.

through their tagging activity. The whole process follows a zoom in & out mode that allows us to (i) have an overall view of the clusters structure and assess the clustering quality (*zoom-out*), and (ii) examine a specific cluster and have an inside look into particular properties of the elements that have been grouped together (*zoom-in*). The main contribution of this paper is summarized as follows:

1. Appropriate data structures that represent the temporal and semantic features of STS objects; the proposed structures address specificities of STS data, such as time-evolving nature and large scale.
2. A similarity measure that captures both temporal and semantic similarities between objects in an STS; the introduced similarity measure may be applied for extracting similarities in any Web 2.0 site that provides textual content (e.g. blog posts).
3. A co-clustering approach that examines simultaneously users' and tags' activity patterns, to extract groups of related objects in an STS; the application of a co-clustering method enables correlations between users and tags to be investigated.
4. A series of visualization widgets that allow to explore the zoom-in and zoom-out views on the extracted clusters; the demonstration is done using both real and synthetic datasets.

To the best of authors' knowledge, that is the first attempt in the literature to co-analyze temporal and tag features for mining information in a STS. Therefore, we cannot test the performance of the proposed algorithm and compare it with that of other methods. However, we demonstrate the feasibility of our approach by presenting a proof of concept experiment on a synthetic dataset. Further, zoom in & out views on a real dataset from Flickr serve as evaluation/visualization methods on different aspects of the clusters' structure. These methods are expected to further push the reader's understanding towards the results of this mining process and allow them to gain a qualitative insight in this.

The rest of the paper is organized as follows: The proposed framework is described in Section 2. First, a problem formulation is given and then the mathematical formulations, on which the approach is based, are developed. Experimental results on both real and synthetic datasets are given in Section 3 in a zoom-in & zoom-out mode, as described above. In Section 4 a series of research areas in which the proposed approach may be exploited is presented, along with related work in each area on temporal tag analysis. Finally, we draw our conclusions in Section 5.

2 Time-aware tags and users grouping

Clustering in STSs is often introduced for overcoming the intrinsic limitations these systems have (e.g. tag ambiguity/synonymy, lack of structure, etc) and extracting knowledge from the mass tagging activity (Specia and Motta 2007; Nanopoulos et al. 2009; Zhou et al. 2007; Giannakidou et al. 2008). Most of such existing efforts apply one-way clustering, i.e. either tag or user clustering. But as described earlier, tags and users exhibit tight binding, since, in general, tags reflect users' interests and thus in most cases users' similarity is estimated based upon similarities of their associated tags. Here, we respect this binding by proposing a co-clustering approach, since typically co-clustering facilitates grouping of different datasets elements (Dhillon

2001). Thus, the proposed approach manages to integrate elements of users and tags datasets, which would not be feasible using a typical one-way clustering approach.

Furthermore, the proposed method underlines the dynamic nature of STSs by considering both temporal and semantic aspects of tagging activities. We claim that such groups, exhibiting both semantic and temporal cohesion, can only be extracted via a time-aware clustering method, which will examine tagging activity at varying time scales. Different time scales allow to capture both macroscopic and microscopic aspects of users tagging, where macroscopic views indicate aggregate patterns of the activity during the entire time period, while microscopic views reflect patterns in specific time frames. Each time scale selection reveals a different micro-view of users' interests that affects the current clustering, since the microscopic view on each user/tag is likely to change as the selected time scale (varies) slides across the timeline. For instance, assume there is a user with a regular interest in weddings (i.e. wedding organizer) and an occasional interest in Olympics. Then, by applying the proposed approach and two different time scales, 1 month and 1 year, the user would be assigned in an Olympics-related cluster and in a wedding-related cluster, respectively. A number of applications may exploit such a technique, customized in each case on the topic of user interest (expressed via the tags the users use), e.g. Olympics-related clusters can be exploited by a sports commercial advertising campaign or be embedded in an application, so that users receive personalized Olympics-related news (e.g. announcement of upcoming events), while wedding-related clusters can be forwarded to a wedding magazine to increase its circulation.

The idea of a time-aware clustering approach was addressed in earlier works of the authors (Petridou et al. 2008; Koutsonikola et al. 2009). However those works describe solely user clustering and their objective function cannot be used to capture different timescales. We claim that the co-clustering of users and tags captures more adequately the correlations between these types of objects in an STS and, therefore, leads to a better clustering. We proceed with a problem formulation and a more detailed description of the proposed approach.

2.1 Problem formulation

The dataset contains two basic elements: users and tags that are parts of sets U and T, respectively. We record the users' activity in a set of equal time frames that comprise the entire time-span $\mathfrak{T}$. We assume that this time division results in I time frames of size τ. Then, for each user u_i, we define a vector UT_i that describes the user's time usage distribution as follows:

$$UT_i = [u_{i1}, u_{i2}, \ldots, u_{iI}], \quad i = 1, \ldots, |U| \quad and \quad k = 1, \ldots, I,$$

where u_{ik} is the number of tags user u_i has assigned during time frame k.

Each user u_i, $1 \leq i \leq |U|$, is associated with all the tags s/he has assigned. We use the variable T_{u_i} to describe the tags assigned by user u_i. Given the fact that users express a kind of interest through tagging, T_{u_i} describes the semantics of user u_i interests.

Tags are the second basic element type in our dataset. We use the variable t_j, $1 \leq j \leq |T|$, to denote a tag. Note that each user may have assigned multiple tags and each tag may be assigned by multiple users. Each tag assignment is explicitly associated with a time frame which marks the time the tag assignment occurred. Based on the time frames associated with each tag we can define the time usage distribution for each tag, as follows:

$$TT_j = [t_{j1}, t_{j2}, \ldots, t_{jI}], \quad j = 1, \ldots, |T| \quad \mathit{and} \quad k = 1, \ldots, I,$$

where t_{jk} denotes the number of times tag t_j has been used during time frame k. Table 1 summarizes the basic notation used in this paper.

Given this data, we address the following problem:

Problem 1 (Users-tags co-clustering) Given an STS tagging activity dataset of U users and T tags over a time period $\Im$, we aim at finding K clusters of users and tags that have the same patterns over the underlying time frames.

The first step in determining whether a tag and a user share the same patterns and should, therefore, be assigned in the same cluster is to define the kinds of patterns that are examined. We aim at patterns that capture the meaning of tags, highlight the topics of the users' interest, and represent the temporal dimension of tagging activities. We call the patterns that relate to the meaning of annotations *semantic patterns*, whereas we refer to patterns that depict the temporal aspect of annotations as *temporal patterns*. More specifically,

Semantic patterns are expected to depict the degree of tag usage and relatedness to semantic topics. To detect semantic relatedness between tags and topics, external sources, such as thesauri and vocabularies, can be employed. The semantic patterns of a user u_i are associated with those of a tag t_j, if any of the tags used by user u_i (T_{u_i}) is semantically related to t_j.

Temporal patterns are expected to set the activity related to a tag or a user in time. The temporal patterns of an object can be (i) occasional, i.e. the activity is focused in specific time frames, (ii) regular, i.e. the activity is spread in the entire timespan, or (iii) periodic, i.e. the activity is repeated in regular time frames.

We consider that the users' behavior is tracked by their tagging frequencies at different time frames and the tags usage is captured by their popularity over time and set two critical aspects that define user and tag similarity: (i) time locality, and (ii) semantic similarity. Thus, for a tag and a user to be grouped together, it is a

Table 1 Basic symbols notation

Symbol	Definition
K	Number of clusters
τ	Time frame duration (time scale)
I	Number of time frames
U	Users' set $\{u_1, \ldots, u_{\lvert U\rvert}\}$
T	Tags' set $\{t_1, \ldots, t_{\lvert T\rvert}\}$
$\mathfrak{T}$	Time period, i.e. the set of time frames at a given τ, $\{i_{1\tau}, \ldots, i_{I\tau}\}$
T_{u_i}	Set of tags assigned by user u_i
u_{ik}	Number of tags assigned by user u_i during time frame k
t_{jk}	Number of times the tag t_j was assigned during time frame k
UT	Two dimensional $\lvert U\rvert \times I$ matrix with elements u_{ik}
TT	Two dimensional $\lvert T\rvert \times I$ matrix with elements t_{jk}

prerequisite that both the user and tag exhibit similar semantic and temporal patterns at simultaneous time frames, as those are defined above. In an effort to capture clusters at different time scales, we use variant size time frames.

2.2 Time-aware co-clustering algorithm

Considering the large number of users and tags in a STS and to proceed to a feasible clustering, we propose spectral-oriented clustering techniques, which perform dimensionality reduction. For that reason we have, also, chosen fast similarity measures (inner product), and fixed data structures (similarity matrices). Figure 1 illustrates an overview of our proposed approach, which is described next in a step-by-step fashion.

Step 1: Preprocessing

Initially, some pre-processing takes place that aims at the finalization of U and T datasets. Many STSs allow only single-word tags and many users have adopted this tags-to-one-word tagging practice, resulting in many useless compound terms. During this step, we analyze such terms and decompose them to their constituent tags. To do so, we employ stemming techniques (Porter 1997) and the lexicon WordNet (Fellbaum 1998). Additionally, during preprocessing we remove rare elements (i.e. users and tags with very low activity), since typically such objects are considered noise in STS analysis.

Step 2: Capturing temporal locality

To capture temporal patterns of a user u_i or a tag t_j for a specific timescale (i.e. time frame duration τ), we use the corresponding vectors UT_i and TT_j. As mentioned in Section 2.1, these vectors represent user tagging activity and tag popularity, respectively, at each time frame. Then, to compute the temporal locality $TemSim$ between the user u_i and the tag t_j, a vector similarity function such as the inner product can be used, as follows:

$$TemSim(u_i, t_j) = \frac{\sum_{k=1}^{I} u_{ik} \cdot t_{jk}}{\sqrt{\sum_{k=1}^{I} u_{ik}^2 \cdot \sum_{k=1}^{I} t_{jk}^2}}$$

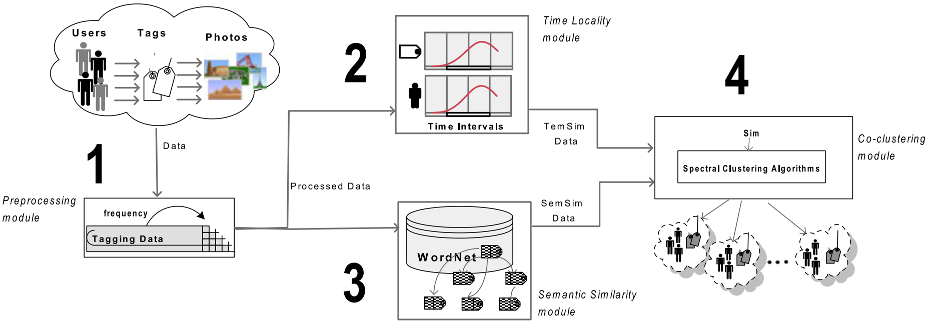


Fig. 1 Time-aware tag/user co-clustering overview

It is worth noticing that choosing a different τ results in different UT_i and TT_j vectors and, thus, in different time locality values.

Step 3: Capturing semantic similarity

To capture the semantic closeness between a user u_i and a tag t_j, we assume that each user u_i is represented by the tags in T_{u_i}, i.e. the tags the user u_i has used. Thus, we actually reduce the user-tag semantic similarity problem to a tag-tag semantic similarity problem between t_j and the tags that belong to T_{u_i}. Intuitively, we choose the maximum similarity value, to ensure that if the user u_i has used the tag t_j or another one semantically close to t_j, then the similarity between u_i and t_j will be maximized. More formally, we define the semantic similarity, *SemSim* between u_i and t_j, as:

$$SemSim(u_i, t_j) = \max_k similarity(t_k, t_j), \quad \forall t_k \in T_{u_i},$$

where the *similarity* function between two tags may be calculated according to any known text similarity metric, such as the *Wu & Palmer* metric (Wu and Palmer 1994) (used in this work), a WordNet-based measure, whose application on tag data is simple and straightforward (Angeletou et al. 2008). The choice of the proposed metric does not affect significantly the performance of the algorithm; any other metric that calculates semantic similarity between two terms would have been appropriate. We chose to employ a WordNet-based solution, since the idea of using knowledge sources, such as WordNet, Wikipedia, to calculate tag similarity has been applied successfully in STS data analysis (Angeletou et al. 2008; Sigurbjornsson and van Zwol 2008).

Step 4: Co-clustering users and tags

To ensure that users and tags assigned in each cluster are related to the same subject topics at similar time intervals, we need to capture similarities both in their temporal and semantic patterns. To this end, we need to define a similarity measure that considers jointly temporal locality and semantic similarity in STSs activities and then apply the co-clustering approach in a binding way between users and tags. Considering the usually huge size of the STS dataset, it is important to look for a measure that performs such calculation, while keeping relatively low computational complexity. To this end, we use the dot product of *SemSim* and *TemSim* vectors, whose computation has $O(n)$ complexity:

$$Sim = TemSim \bullet SemSim$$

Thus, we construct the **Sim** matrix that can be said that represents the user tagging activity in an STS in a vector form, where rows denote users, columns denote tags and each element *Sim*(*i*, *j*) expresses the semantic and temporal similarity between the tags assigned by user *i* and the tag *j*.

Given **Sim**, we may proceed with the application of the co-clustering algorithm (Dhillon 2001), in order to get clusters containing users and tags with similar patterns over time. The applied algorithm is based on the spectral clustering theory, as discussed in Giannakidou et al. (2008) and Koutsonikola et al. (2008), and relies on the eigenstructure of the similarity matrix, **Sim**, to partition users and tags into *K* disjoint clusters. Specifically, our spectral partitioning method uses the left and right singular vectors of an appropriately scaled matrix **Z** that yields as follows:

First we calculate the diagonal matrices $\mathbf{D_u}$ and $\mathbf{D_t}$, such that:

$$D_u(i,i) = \sum_{j=1}^{|T|} Sim(i,j) \quad \text{(Sum of tag weights incident to the user } u_i\text{)}$$

$$D_t(j,j) = \sum_{i=1}^{|U|} Sim(i,j) \quad \text{(Sum of user weights incident to the tag } t_j\text{)}$$

Then, we perform a singular value decomposition on the matrix $\mathbf{NSim} = \mathbf{D_u^{-1/2} Sim D_t^{-1/2}}$ and obtain the $l = \lceil \log_2 K \rceil$ left and right singular vectors lsv_2, lsv_3, $\ldots lsv_{l+1}$ and rsv_2, rsv_3, $\ldots rsv_{l+1}$, respectively. These singular vectors are known to contain K-modal information about the dataset in discourse (Dhillon 2001). Next, we define the matrix **Z** as:

$$\mathbf{Z} = \begin{vmatrix} D_u^{-1/2} LSV \\ D_t^{-1/2} RSV \end{vmatrix}$$

where $\mathbf{LSV} = [lsv_2, lsv_3, \ldots lsv_{l+1}]$ and $\mathbf{RSV} = [rsv_2, rsv_3, \ldots rsv_{l+1}]$. Finally, we apply on the reduced-dimensionality matrix **Z**, a *K-means* algorithm and obtain the desired K-partitioning of both users and tags.

3 Zooming in & out on time-aware clusters

To evaluate the proposed algorithm we carried out experimentation on both synthetic and real datasets. In this section, a cluster-analysis is presented that involves clusters' visualization at a focused (zoom-in) view and at an overall (zoom-out) view, in order to highlight the contribution of the proposed clustering scheme. More specifically, we experimented with synthetic datasets of different sizes (which were generated as described in the next section) to examine whether the proposed algorithm captures the underlying data structure. Then, we tested our method on a Flickr dataset with photos and their associated metadata (i.e. tags, uploading time, user, etc), which were uploaded during the time period September 2007–September 2008. After the preprocessing step, we resulted in a dataset of 1,218 users, 6,764 photos and 2,496 unique tags that span in 210 days. The input parameters used are the cluster number K and the time frame duration τ.

3.1 Zooming-in to synthetic dataset clusters

For the purpose of this experimentation, we generate a series of datasets that contain rough semantic descriptions and temporal features of activities in an STS. Data generation was based on a specific model, so that we can test whether the co-clustering algorithm succeeds in discovering that model. More specifically, as discussed earlier, the idea in the proposed co-clustering approach is to capture user groups dealing with specific topics for specific time periods and associate them with tag groups that describe the same topics in the same time periods. In case, we detect similar semantic and temporal patterns in these groups (cf. Fig. 2), then they are

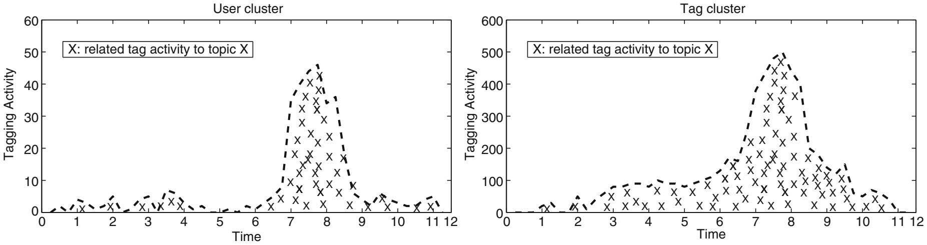


Fig. 2 A user and a tag cluster with similar semantic and temporal patterns. The user cluster contains users dealing with topic X in a specific time period, while the tag cluster contains tags related to topic X and assigned, approximately, in the same time period

assigned in one cluster. Towards this end, the datasets are generated in such a way so that there are such groups. It, therefore, remains to be seen if in fact these groups are detected by the proposed method and assigned in the same cluster.[2]

To address this requirement, we created an initial synthetic dataset that contains clusters of users and tags with similar temporal and semantic patterns, as shown in Fig. 3. Specifically, the synthetic dataset contains four user clusters and three tag clusters, according to their so-called temporal locality. Moreover, the dataset is divided in four clusters, based on the semantic similarity of objects. As can be easily drawn from the Fig. 3, there are four clusters with similar temporal and semantic patterns that should be extracted from the proposed approach. The description of the dataset generation process together with the application of the proposed co-clustering algorithm follow:

We consider the set of users U and the set of tags T that exhibit activity in an STS over a period $\mathfrak{T}$ of I time frames. We assume that users are divided in advance into $K = 4$ clusters according to their time locality i.e. the way they perform their tagging activity over time. In other words, users of the same cluster assign approximately a similar number of tags in the same time periods, for instance the users belonging to the first cluster are very active during the first two time frames, the ones belonging to second cluster are very active during the last time frames and so on. In order to produce such a cluster structure, we repeat for each cluster the following steps:

- Define a random number of members (i.e. users).
- Select for each time frame a mean value $\mu_{i,j}$, which is uniformly distributed in [0...99].
- Generate points in each time frame, by adding values sampled from the normal distribution $N(\mu_i, \sigma^2)$. Each point in each time frame corresponds to a user's activity in the specific time frame, thus, the total number of points in each time frame equals the number of the cluster's users.

Having completed the described process for the four clusters, we form the users over time matrix (the $|U| \times I$ **UT** matrix) in a way that there is an accordance among users of the same cluster on the time they perform their tagging activity. Then, we

[2] The potential applications of tracking such clusters are discussed in the next section.

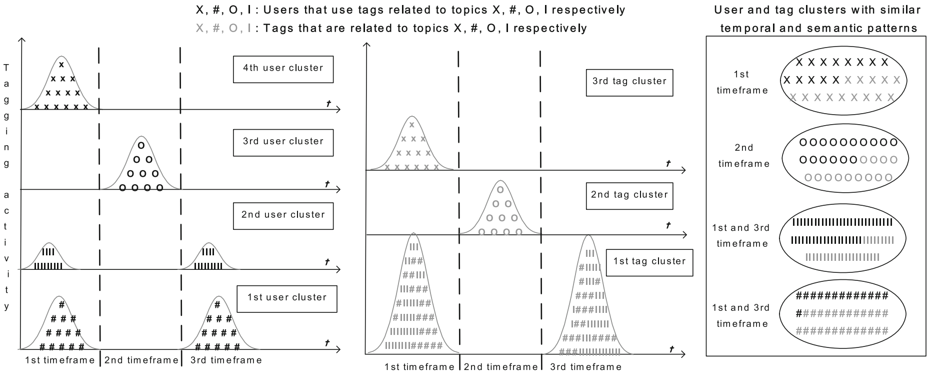


Fig. 3 Semantic and temporal patterns on user and tag clusters generated in the synthetic dataset. The particular dataset is generated in such a way, so that tag clusters are not well separated neither in temporal dimension nor in the semantics and it is the users' activity that drives the clustering process and enables the tracking of four clusters with similar semantic and temporal patterns

create the tags over time matrix (the $|T| \times I$ **TT** matrix), by following exactly the same process for $K = 3$. Thus, we obtain three sets of tags, each of which contains tags that are assigned during the same time periods.

According to our scenario, there is 1:1 relationship between clusters of users and tags' topics, i.e. we consider that the set of T tags refers to four different topics. In other words, users belonging to the same cluster are highly interested in one of the four topics and show less or no interest for the rest three of them. Thus, a set of tags that belong to a topic are randomly assigned to users who are members of the same cluster. A small percentage of different topic tags is also assigned to the aforementioned users.[3] For the purpose of our experiments we assume that the users of two of the four clusters, even though they are interested in different topics, they perform some of their tagging activity concurrently. This explains the generation of three clusters in **TT** matrix, as described earlier.

For our experiments we fixed the values of users to be $|U| = 600$, $|T| = 2{,}000$ and $I = 60$ time frames, while we set the value of standard deviation to $\sigma = 1$. In order to efficiently depict our dataset properties we proceed to a zoom-in view to identify and elaborate on existing specificities. To this end, we employ advanced multivariate graphical techniques such as Andrews' curves (Andrews 1972), which are suitable in case of high-dimensional data. Andrews' curves is a way to visualize and hence to find structure of high-dimensional data. Each multivariate observation e.g. $(UT(i,1), \ldots UT(i,I))$ is transformed into a curve based on the function:

$$f(z) = UT(i,1)/\sqrt{2} + UT(i,2) \cdot \sin(z) + UT(i,3) \cdot \cos(z) \\ + UT(i,4) \cdot \sin(2z) + UT(i,5)\cos(2z) + \ldots$$

and plotted over the range $-\pi \leq z \leq \pi$. Thus, each data point (e.g. user, tag) may be viewed as a curve between $-\pi$ and π. This function representation has

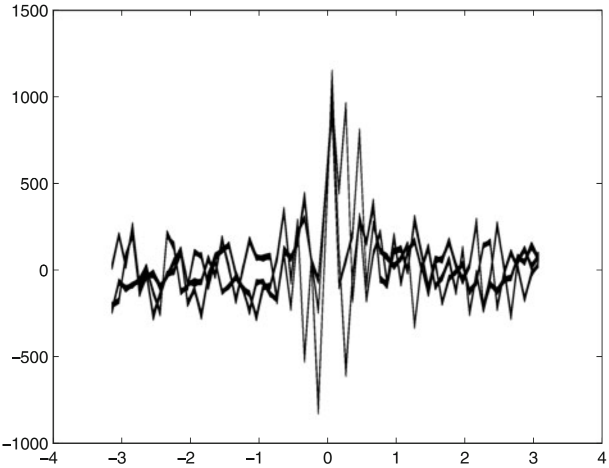
[3] This is not imprinted in Fig. 3 for clarity reasons.

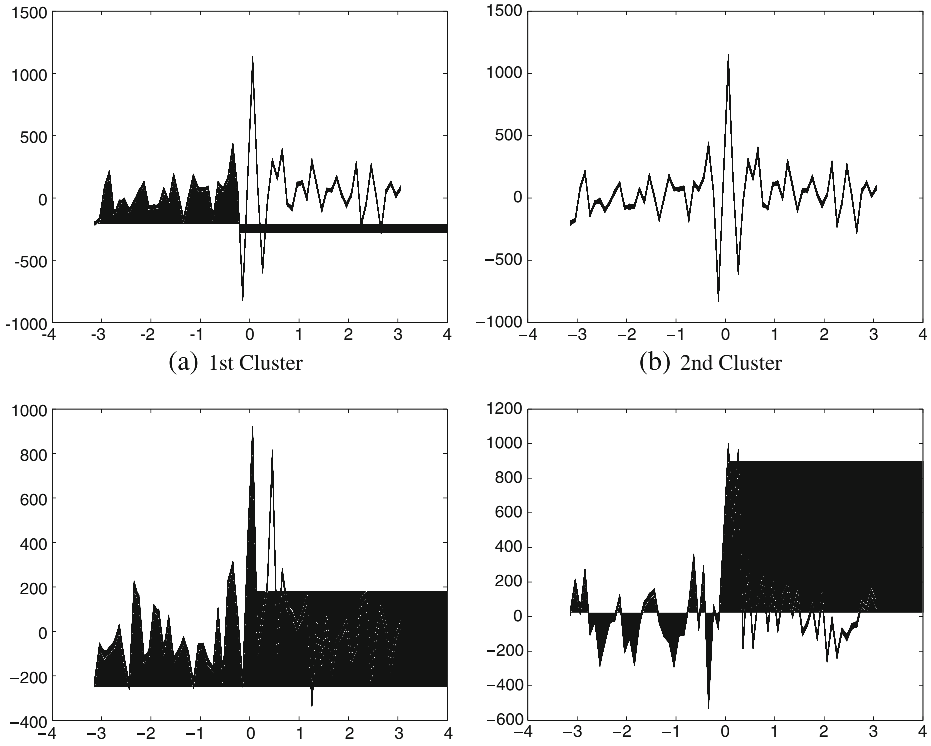
several interesting characteristics, namely it preserves the standard deviation and the distances of data points (e.g. close points will appear as close curves while distant points as distant curves). So, if there is an underlying structure in the data, it may be visible in its Andrews' curves. More specifically, regarding Andrews' curves in conjunction with the clustering process, we can claim that the different shapes of curves among clusters are an indication of dissimilarity between objects belonging to different clusters while the similar curves among objects of the same cluster are an indication of similarity between them (Theodosiou et al. 2007). Figure 4 depicts a zoom-in view in the temporal dimension of the tags of the dataset. It is evident that, based solely on this dimension, there are three clusters of tags over time, since three curve shapes have been detected.

We apply the proposed time-aware co-clustering algorithm in this dataset, in order to see whether our clustering algorithm will manage to identify the four user and tag clusters with similar temporal and semantic patterns, as shown in Fig. 3, although according to the time aspect we can only identify three tag clusters. In other words, we aim to see whether the consideration of both time and semantic aspects enhances the overall clustering process.

At the end of the clustering process we obtain successfully the four user clusters which were indicated by both time and semantic aspects. However, the interest in our case is located on the tag clusters formation because of the different feedback provided by the time and the semantic similarity between the tags in the dataset. The clustering results with respect to tags are depicted in Figs. 5 and 6. The proposed algorithm resulted in four clusters of tags and therefore, it managed to successfully identify the four different clusters, even though information based on the time locality indicated that there were only three clusters. Figure 5 presents these clusters in terms of time locality, while Fig. 6 presents their semantic aspect. Specifically, each subfigure of Fig. 5 presents one of the obtained clusters. It is evident that Fig. 5a, c and d depict three different shapes of curves which correspond to the three tag clusters which have been initially generated in terms of time. Figure 5a and b present the same shapes of curves, containing, thus, tags that show similar activity in time. The proposed algorithm however was not misled and managed to

Fig. 4 Tags over time Andrews Curves





(a) 1st Cluster

(b) 2nd Cluster

(c) 3rd Cluster

(d) 4th Cluster

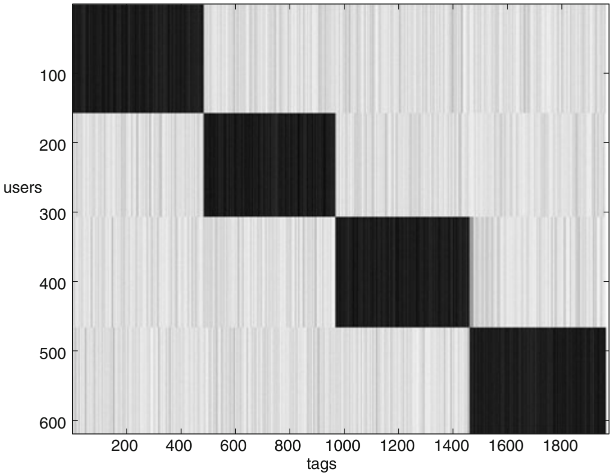
Fig. 5 Tags clusters over time

successfully separate the tags of these two clusters. It is the semantic aspect of tags which is highlighted by users preferences and imposes their separation into different clusters.

In terms of the compactness of the obtained clusters, a visualization that allows such a view is depicted in Fig. 6 where the similarities between users and tags of all clusters are graphically represented. Rows correspond to users and columns to tags and they have been rearranged so that users and tags of the same cluster are put in consecutive rows (columns). Moreover, the darker the coloring of a cell (i, j), where $1 \leq i \leq |U|$ and $1 \leq j \leq |T|$, the more close in terms of the similarity *Sim* the corresponding user and tag are. Thus, given that clusters contain the most similar users and tags, the darker rectangles appear on the plot's diagonal and reveal the clusters of our dataset. Furthermore, the intense color difference between points in the diagonal and outside the diagonal indicates that users and tags that have been assigned to the same cluster present strong similarity and therefore clusters are significantly coherent.

We have also performed the clustering based only on the *TemSim* similarity measure, in order to examine the clustering results in case we considered only the time aspect. The algorithm resulted in four clusters of users and three clusters of tags. Thus, it assigned tags that referred to two different topics to the same cluster

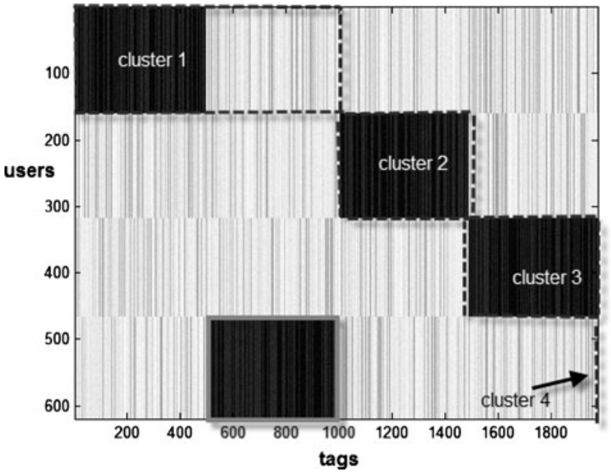
Fig. 6 Semantic aspect of clusters



because their assignment was performed by users the same time periods. Figure 7 presents the obtained clusters in terms of their semantic similarity. Clusters are denoted with the dotted outline. Since there is no tag assigned to the fourth cluster, the respective rectangular in the diagonal has no width (it contains only users). Moreover, cluster 1 contains a set of tags (the first, dark, half rectangle) which present significant semantic similarity with users of the cluster and a set of tags (the second, light-colored, half rectangle) which is not semantically similar to the clusters' users. Nevertheless, the tags of the second half of the first cluster present high similarity with the users of the fourth cluster. Thus they should have been assigned to the fourth cluster which contains no tags. In that case we would have obtained the results of Fig. 6.

We have also experimented with another case scenario where there were more user and tag clusters with respect to the temporal dimension, than by considering

Fig. 7 Semantic aspect of clusters considering time locality



only the semantic aspect (i.e. the number of generated clusters in the **UT** and **TT** matrices were larger than the number of topics in the dataset). Following a similar evaluation process we saw that time guided the clustering process in more accurate and detailed in separation clusters since clusters are identified based on more criteria. It is therefore evident that considering both the time and semantic aspects guides the clustering process in more enriched with information and coherent clustering results.

3.2 Zooming-out on synthetic dataset clusters

We proceed to an analysis that highlights the overall performance of the proposed approach and offers a global view—that is a *zoom-out view*—on the extracted clusters. To evaluate the scalability of the proposed clustering approach we have generated datasets of different size. Furthermore, in order to study the algorithm's performance on more loose associations between users and tags in an STS, we experimented with various values of standard deviation σ, which denotes the degree of separation between cluster members. All the datasets presented here were generated based on the same model our initial dataset was created (cf. Fig. 3). For the evaluation of the obtained clusters we have employed the F-measure which combines the ideas of precision and recall and it is a broadly accepted and reliable index used in various clustering evaluation approaches (Larsen and Aone 1999). Precision and recall express the degree of relevancy of the data points (users or tags) assigned to clusters and they are defined according to the following equations:

$$precision = \frac{|\{relevant\ points\} \bigcap \{assigned\ points\}|}{|\{assigned\ points\}|}$$

$$recall = \frac{|\{relevant\ points\} \bigcap \{assigned\ points\}|}{|\{relevant\ points\}|}$$

Given the precision and recall definitions the F-measure is defined as:

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

F-measure fluctuates in the interval $[0, \ldots, 1]$ with higher values indicating a better clustering scheme.

Table 2 presents the F-measure values for different datasets and standard deviation set to $\sigma = 1$. The high F-measure values show that the proposed clustering scheme manages to retain its good performance in case larger size datasets were employed. As we can see, in two of the four clusters there are some data points that were not successfully assigned. This is due to the fact that according to temporal

Table 2 F-measure values for different size datasets

Users	Tags	Time frames	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1,594	3,613	233	0.98	0.97	1	1
1,616	3,576	200	0.98	0.99	1	1
2,133	6,787	200	0.99	0.99	1	1
4,144	13,181	200	0.99	0.97	1	1

Table 3 *F*-measure values as a function of standard deviation

Standard deviation	Cluster 1	Cluster 2	Cluster 3	Cluster 4
$\sigma = 0.5$	0.99	0.99	1	1
$\sigma = 1$	0.95	0.99	1	1
$\sigma = 1.5$	0.93	0.96	0.98	1
$\sigma = 2$	0.93	0.92	0.97	1

aspect the two clusters should be considered as one. However, these mis-assignments represent a very small fraction of the overall dataset.

Finally, as in the initially generated dataset we chose a model that minimized the inter/intra-cluster scatter, we, also, evaluated the performance of the proposed clustering scheme for different values of standard deviation in order to examine whether its performance is affected in case separation between clusters becomes harder detectable. The values of *F*-measure for a dataset consisting of 1,400 users, 4,500 tags and 60 time frames as a function of standard deviation are depicted in Table 3. The results prove that the proposed algorithm manages to reveal the dataset's underlying structure even in cases of high standard deviation values.

3.3 Zooming-in to real dataset clusters

Furthermore, we tested our approach on the Flickr dataset that consists of 1,218 users, 2,496 unique tags and 6,794 photos. The collected dataset refers to 210 distinct days that span the period from September 2007 to September 2008. The selection of the data was based on the following four seed tags: `ancient greece`, `earthquake`, `wedding`, `olympics`. The particular topics were selected intentionally, since they are associated with world-class events that occurred at the specific period (i.e. Olympic games 2008, 2008 Sichuan earthquake in China) and we are interested in examining whether the proposed time-related analysis is able to track such incidents. Furthermore, we chose topics like "ancient greece" and "Olympics", since the tags describing these topics are semantically related, and we wanted to test whether the proposed method will manage to distinguish between regularities and irregularities in user interests (i.e. irregularities usually occur if users follow events or trends and they are imprinted as sudden bursts of tagging activity at specific time periods). In our case, the topic of Olympics shows an occasional interest, whereas the topic of ancient greece a regular interest.

At the end of the time-aware co-clustering process on the Flickr dataset, each cluster holds temporal and semantic patterns that describe the topic of the tagging activity inside the cluster and the time frames this activity occurred. To gain insights into such patterns, we zoom-in to one particular cluster and study the tag usage, using a clock-like visualization metaphor. First, we gather the tags that have been assigned to at least 60% resources of the specified cluster, which constitute the *cluster topic*. Then, we visualize the *cluster topic* along with the associated time frames as a "clock", in the following fashion: We indicatively set $\tau = 10$, which results in 21 time frames[4] being depicted on the clock's periphery. Each tag is rendered as an arrow stemmed

[4] Each time frame corresponds to 10 days duration, at the particular time scale.

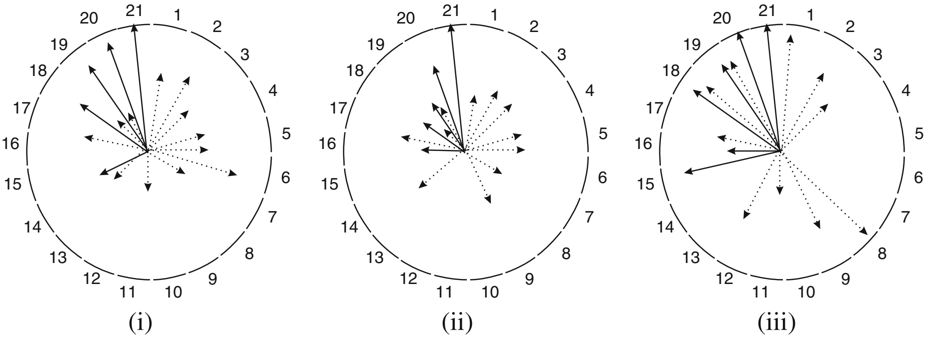


Fig. 8 Clock-like visualization of tags (*i*) and users (*ii*, *iii*) patterns in a cluster

from the circle's center. The length of the arrow indicates how frequently this tag was used, and the orientation of the arrow indicates the time frame that this tag was used. We use a normalization to determine the arrows' length, by dividing with the maximum tag frequency, so that all arrows have length in [0, 1].

Then, each user is also represented as a "clock" by analyzing the tags that s/he has used and the timestamps that the user's tagging activity occurred, in the same vein, as described above. Figure 8 shows indicative clocks of the cluster topic tags (i) and two random users (ii, iii) of this cluster. The emergent cluster topic constitutes of `ancient-greece` related tags (shown with dotted lines) and `Olympics` related tags (shown with solid lines). We can see that the `ancient-greece` related tags spread over the entire time span, while the `Olympics` related tags fall mostly in the last five time frames (August-September 2008). Likewise, the selected users seem to regularly use `ancient-greece` related tags and have an abrupt rise in the usage of `Olympics` related tags during the last timestamps. This massive tagging preference during the last time frames implies that a related event occurred at that period and attracted Olympics-friends to comment on it, through tags. Indeed, this period of burst olympics-related tagging activity coincides with Olympics 2008 at Beijing. Moreover, the described clock-like visualization enables us to distinguish between users' regular and irregular interests (i.e. ancient-greece versus Olympics).

Such visualization patterns reflect the tagging activity in each extracted cluster and, in a well-defined clustering, the objects contained in one cluster (tags and users) emerge alike clocking patterns, as it is shown in Fig. 8.

To highlight the impact of the temporal dimension in the approach, and since it is difficult to examine it for all the objects (i.e. users and tags) due to the dataset's big size, we zoom-in to three randomly chosen users who have relevant topic tagging activity and examine their clustering assignments for $K = 7$ and $\tau = 1$,[5] using either a time-aware clustering approach or a typical (static) tag-based clustering. Figure 9 shows these users' tagging frequency ($axisY$) in the underlying time frames ($axisX$), and their respective tag clouds. All three users refer to `earthquake`-related

[5] Each time frame corresponds to 1 day duration, at the particular time scale.

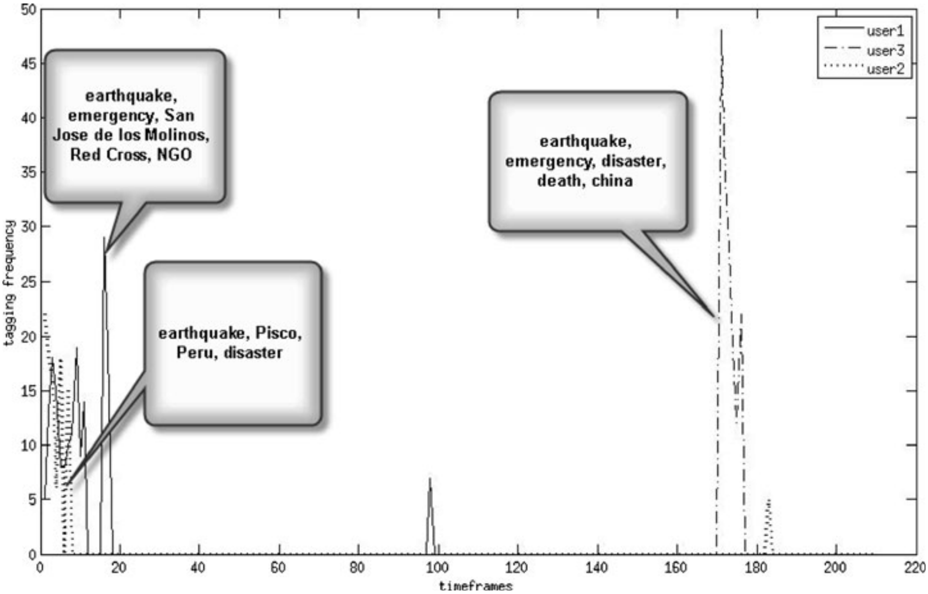
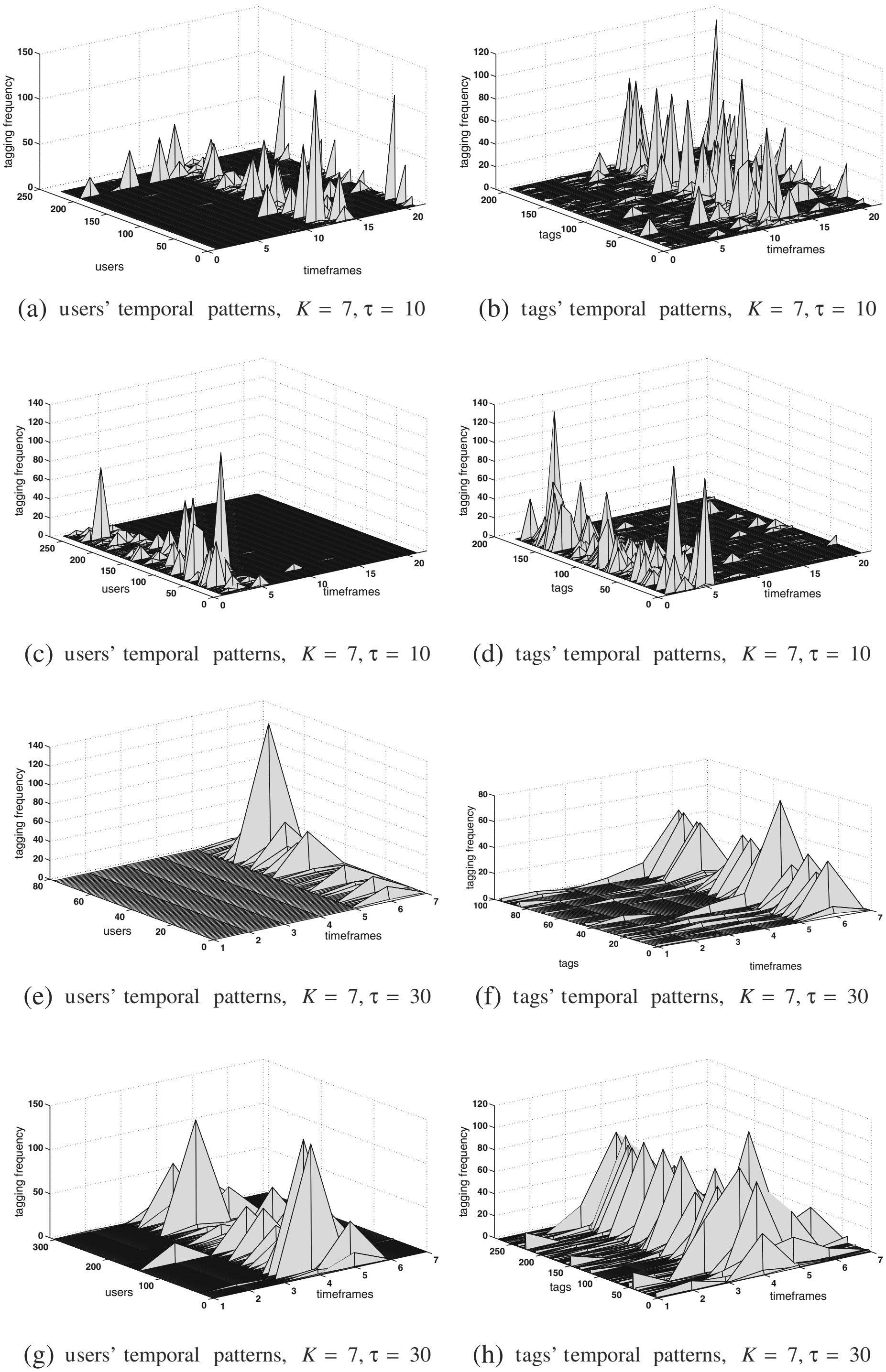


Fig. 9 Zoom in on the impact of temporal aspect in the clustering ($\tau = 1$)

tags. However, user1 and user2 assign tags mainly on the initial time frames that correspond to the period August 2007, while the tagging activity of user3 is recorded during the timeframes that span May and June 2008. Due to the big overlap in their macroscopic tagging activity, all three users are grouped together by a static tag-based clustering approach, even though they differ in terms of the actual time the tagging occurred in each case. On the other hand, the proposed time-aware approach manages to separate these three users and assign user1 and user2 together in one group, and user3 in a different group. It is apparent that user1 and user2 are more similar to each other, since they use related tags on common timeframes. These timeframes cover the period immediately after the Peru earthquake in August 2007, suggesting that user1 and user2 were, probably, attracted by this event, and their interest was occasional, as it lasted only for the time period around this particular event. Thus, user1 and user2 together with other users who assigned similar tags on this time period are grouped in one cluster that can be regarded as the "Peru-earthquake-2007" cluster. On the other hand, user3 seems to have a similar occasional tagging activity on earthquake matters, as he uses tags, such as `earthquake`, `emergency`, `shelter`. However, user3's tagging falls in a different period, that coincides with the 2007 Sichuan earthquake in China, highlighting that this interest was aroused due to a different event, so user3 is well not grouped with user1 and user2.

3.4 Zooming-out on real dataset clusters

Next, we proceed to an overall analysis of the extracted clusters' structure, that is a so-called zoom-out visualization on the obtained clusters. This kind of analysis demonstrates, also, that the proposed approach is beneficial to capturing events



(a) users' temporal patterns, $K = 7, \tau = 10$

(b) tags' temporal patterns, $K = 7, \tau = 10$

(c) users' temporal patterns, $K = 7, \tau = 10$

(d) tags' temporal patterns, $K = 7, \tau = 10$

(e) users' temporal patterns, $K = 7, \tau = 30$

(f) tags' temporal patterns, $K = 7, \tau = 30$

(g) users' temporal patterns, $K = 7, \tau = 30$

(h) tags' temporal patterns, $K = 7, \tau = 30$

Fig. 10 Snapshots of tags and users patterns over time in four indicative clusters (each row represents a cluster)

Table 4 Indicative clusters and emergent cluster topics

Olympics cluster	basketball, climbing, medalist, wrestling, classicalworld, olympicceremony, volunteer, ancientolympia, london2012, athensolympics, peking, butterfly, softball
Wedding cluster	happiness, party, married, groom, bouquet, balloon, wife, husband, romance, love, dancing, batchelormansion, weddingcake, marry, romantic, married

and emergent topics/trends. We begin by visualizing the temporal patterns of the extracted clusters. Indeed, the time-awareness of our method is best confirmed by a macroscopic view of the time—activity relation of the objects contained in each cluster. Figure 10 zooms-out on some indicative clusters, where it is shown that the activity distribution of users and tags assigned in each cluster is similar over time. More specifically, each row depicts a cluster, where in the left (right) the users' (tags') activity is shown as a function of their tagging frequency (*axis Y*) and the time frames (*axis Z*). It can be seen that the proposed method achieves to group together users and tags that are active during the same time periods. Moreover, Fig. 10 illustrates different clustering for varying values of τ, highlighting, thus, that the proposed method can be used to obtain clusters that expose user choices at particular time frames. Such a visual analysis reveals the temporal patterns that underlie each extracted cluster, i.e.: *occasional*, in which the activity is focused only around specific time frames (e.g. Fig. 10c and d, e and f), *regular*, where the activity is spread in the entire timespan (e.g. Fig. 10a and b, g and h), and, *periodic*, in which the activity occurs in repeated patterns. Occasional and periodic patterns can be exploited for event recognition, as discussed in the Section 4.

Each cluster's tags reflect the interests of the included users at particular time frames, that is the semantic patterns of each cluster. By examining the tag clusters and selecting the most frequently used tags we can map the tagging activity to meaningful topics. Table 4 zooms out on the most prominent tags of two clusters, a regular and an occasional one (whose temporal patterns are depicted in Figure 10a, b and e, f, respectively). Specifically, the first cluster constitutes the semantic zoom-out of the occasional "Olympics2008-friends" cluster, while the second one corresponds to a "wedding-related" cluster that includes users with a regular interest in wedding, e.g. wedding organizers.

4 Time-aware clustering in practice and related work

Temporal analysis has been an active topic of research in many fields. A detailed study on examining temporal dynamics on online data, in general, can be found in Kleinberg (2006). In the present article, a socio-temporal approach was presented that exploits temporal patterns, considering not only the temporal aspect, but also the interactions between the components in an STS (users, tags, resources) and how these interactions evolve over time. Observing the zoom in and out views, in the previous section, we can infer that the proposed clustering approach yields temporal and semantic patterns of users and tags that can be exploited in a number of research areas. Below we quote some of these fields, along with related work that involves temporal tag analysis in each area.

Topic detection, tracking and trend analysis locate coherent topics out of unstructured sets of tags in an STS and identify "hot" topics that signify emerging trends. The common ground in approaches that perform this kind of analysis is that they divide the entire time period under consideration into subsequent time frames and, afterwards, they calculate a metric's value in each time frame that attempts to capture a tag's significance in the specified time frame. By employing statistical methods and analysis over time of the tags' values, they detect "anomalies" or bursts in the tag distributions that are interpreted as *topics of interest* or *trends* for the given time periods.

More specifically, Sun and colleagues (Sun et al. 2008) use the χ^2 statistical model (Swan and Allan 1999), to determine whether the appearance of tag t in a time frame i is significant and, thus, to discover tags that constitute "topics of interest" at particular time frames. They apply the model on a 2×2 contingency table, which records all tagging activities that include and not include tag t in a time frame i and in time frames $i_{\text{past}} < i$. Their analysis uses one degree of freedom and highlights as trends the tags whose χ^2 value reports a probability of chance $\preceq 0.005$ on one time frame. Wetzker and colleagues in Wetzker et al. (2008a) claim that a tag signifies a trend, if it attracts significantly more new users in a currently monitored time frame than in past time frames. In their analysis, they use a probabilistic generative model to describe tag distributions which result from the tags' counts at two consecutive time frames. Their tag popularity metric contains a parameter ϵ that represents the assumed prior frequency of each tag and it allows to distinguish between relative and absolute increase of a tag's popularity. Further, their technique differentiates in that they apply smoothing to cope with the problem of data sparsity and in that they use the *diffusion-of-attention* concept (Wetzker et al. 2008b), to reduce the effect of spam. A trend detection measure is introduced in Hotho et al. (2006a), which captures topic-specific trends at each time frame and is based on the weight-spreading ranking of the PageRank algorithm (Brin and Page 1998). More specifically, the so-called FolkRank measure ranks the closeness of each tag to a topic with respect to its importance in relation to a given preference vector (Hotho et al. 2006b).

Such techniques are useful especially in the marketing domain for business planning, policy making and so on. The proposed approach may be aligned with the research in this area by examining the semantic patterns of each extracted cluster and considering them as topics of interests of particular users for the time periods analyzed. Furthermore, the observation of occasional temporal patterns in clusters may be interpreted as a trend detection method that captures users' following certain topics over specific time periods.

Event recognition analyze tags/time usage patterns along with geo-related information available in an STS and identify real-world events. The topic of time-based event recognition has been investigated in the literature under the *Topic Detection and Tracking* area, (TDT), described earlier (Allan 2002). However, here we present it separately, to quote the challenges involved and illustrate the kind of analysis that can take place, regarding event recognition in the STS context.

Rattenbury and colleagues (2007) infer event semantics of tags in an STS, by examining the tags' distribution over time and space. The intuition behind their method is that a tag describing an event usually occurs at a specific segment of time and is assigned on photos geo-tagged around a specific place. (e.g. "olympics2008"). In their proposed method, they introduce an approach that does not rely on a-priori defined time frames, but searches for low-entropy clusters in the time usage distribution of a tag that are robust at many time scales. Thus, they manage to capture events of different time scales, which constitutes a major research challenge met in this area. Further, they tackle the issue of periodicity by checking a set of constraints and applying the modulo function, so that clusters that are apart the same distance from one another along the temporal dimension can be treated as a single cluster (i.e. event). Other methods that can be employed, in order to search for tags that can be mapped to events are standard burst detection techniques, such as *Naive Scan Methods* (Vlachos et al. 2004) or *Spatial Scan Methods* (Kulldorff 1999), borrowed from other domains. The latter, though, rely on a-priori defined time frames, and perform worse. Becker and colleagues (2010) present similarity metrics to mine clusters of social media content associated with real world events.[6] They rely on combining multiple context features that are inherent to this type of content (i.e. tags, textual features, time, location) and implementing similarity metric learning algorithms. The performance of the proposed algorithms was analyzed on real event-related datasets from Flickr and the relative contribution of each feature on each metric was estimated. The temporal proximity between objects was considered in all metrics and, in most cases, was ranked as an important feature. Further, in order to tackle scalability issues that arise due to the large volume of data in STSs, they employ incremental clustering algorithms and model the notion of similarity between an STS object and the centroid of a cluster of STS objects.

Such techniques are based on the fact that most people are fond of sharing their own contributed content in STSs. Therefore, the potential that large collections of such content associated with particular events could be mined from these systems increases. An example application that is enabled from this kind of analysis is the integration of faceted browsing of events and related activities in browsers. Although in the proposed time-aware co-clustering approach no spatial features are considered, the analysis of clusters on the real Flickr dataset showed that the method succeeded in tracking events. As mentioned the dataset was selected in such a way, so that photos of two world-class events, namely *Olympics 2008* and *Sichuan 2008* earthquake, were included. Having repeated the proposed clustering process for various timescales (i.e. various values of τ), we always resulted in clusters referring to each particular event exclusively. Moreover, clusters that refer to events of more restricted scale were identified, but not described here due to space restriction. To track events in the proposed method, we focus on clusters with occasional or periodic temporal patterns.

Information visualization illustrate tagging activity in an STS with an explicit temporal dimension. Applications that aim at information visualization in an STS should combine back-end analysis for capturing interesting tags at given time periods and front-end analysis for illustrating this information to the user and, at the same time, allowing the user to interact with the system (for example, the user may decide

[6] Although the approach in Becker et al. (2010) is more related to our approach than the other ones presented, still there cannot be a direct comparison between the two methods, since the one mines resources, whereas the other groups together tags and users.

to view information on a different time scale). Regarding the backend analysis, the algorithms described in TDT can be employed. Especially, for the task of front-end analysis the use of visual metaphors facilitates user experience in such applications.

Dubinko and colleagues (2006) developed a browser-based application in which the user may navigate through interesting tags of various time frames in Flickr, at varying timescales. They grasp a tag's interestingness on a particular time frame by counting its frequency in this time frame over other time frames. In order to achieve efficiency, they employ backend algorithms that pre-compute tag interestingness scores for varying sized time frames. Russell (2006) presented a tool that visualizes the collective tagging activity on a resource over time, highlighting periods of stable and changing tagging patterns. The latter denote a change in users' awareness of the described resource.

Here, a number of visualization widgets have been proposed that allow the reader to explore temporal and semantic patterns out of the user tagging activity in an STS.

Furthermore, the simultaneous integration of users and tags in one cluster that was proposed here renders a number of user centric scenarios in the following areas:

Community-based tag recommendation support each user in the tagging process by suggesting tags that are used by users in the same cluster. The use of recommendations is often motivated in STSs, since they facilitate the task of tagging and promote vocabulary convergence (Hotho et al. 2006a). The presented co-clustering approach results in user communities being linked with tag communities and we can claim that each tag group captures the vocabulary of the user community, it is linked with. Thus, a recommender system may integrate the proposed method to support STS users.

Personalization find mechanisms that acquire and represent user interaction data, so as to build profiles that will help better understand users' needs and, therefore, provide individualized services (Shepitsen et al. 2008). The presented method enables extraction of tag clouds for a specific user at various time scales, which can be used to build time-aware (i.e. dynamic) user profiles. Such profiles reflect user interests and habits at different time periods and can serve as input to a personalized application for implementing adaptive behavior.

Fighting spam on STSs detect malicious users' misleading descriptions that aim at attracting user visits, through a ranking mechanism according to which regular users outweigh the occasional ones. As STSs count on user-generated content, they offer a tempting target for spam (Heymann et al. 2007). Our method can be used to generate a user ranking in the following fashion: users that appear to have a regular interest around a topic get a higher rank, where those that seem to be rather occasionally interested in a topic get a lower rank. This way, users build a ranking-based reputation that can serve as contribution quality indicator.

5 Conclusions

The underlying idea in the proposed approach is the integration of temporal information together with a typical tag similarity metric, in order to get users/tag clusters in an STS. The use of co-clustering vs typical one-way clustering yields not only relations between users, but, also, topics of their common interests. The proposed co-clustering approach has been evaluated on both synthetic and real datasets. Zoom in and out views on the extracted clusters reveal semantic and temporal patterns in the dataset that identify periodicity, regularities, topics and trends with regard to an individual user or the entire user community. Beyond this method's effectiveness, we believe that one of the most interesting challenges in this area is the automatic detection of the time frame's duration, τ. Our future work will address how to automatically identify chunks of time that capture the nature of an event or imply an interesting formation between objects in an STS. Specifically, we aim to test the use of some well-known techniques of advanced discretization, such as entropy based discretization (Fayyad and Irani 1993) and statistic-based methodologies for determining discretization intervals (Richeldi and Rossotto 1995). Furthermore, we intend to investigate customizations, so as the discretization method best respects the multidimensional nature and other specificities of STS objects (Wu et al. 2004).

Acknowledgements This work was supported by the FP7 project WeKnowIt, partially funded by the EC under contract number 215453.

References

Allan, J. (2002). Introduction to topic detection and tracking. In *Topic detection and tracking: Event-based information organization* (pp. 1–16). Norwell: Kluwer Academic.

Andrews, D. F. (1972). Plots of high-dimensional data. In *Biometrics* (Vol. 28, pp. 125–136). Alexandria: International Biometric Society.

Angeletou, S., Sabou, M., & Motta, E. (2008). Semantically enriching folksonomies with flor. In *Proceedings of the 5th ESWC workshop: Collective Intelligence and the Semantic Web*.

Becker, H., Naaman, M., Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining* (pp. 291–300). New York: ACM.

Begelman, G., Keller, P., & Smadja, F. (2006). Automated tag clustering: Improving search and exploration in the tag space. In *Proceedings of the collaborative Web tagging workshop, 15th international World Wide Web conference (WWW'06)* (pp. 89–98). Endinburgh, Scotland.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN. Systems, 30*, 107–117.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*. San Francisco: ACM.

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., & Tomkins, A. (2006). Visualizing tags over time. In *Proceedings of the 15th international conference on World Wide Web* (pp. 193–202). Edinburgh: ACM.

Fayyad, U. M., Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI'93* (pp. 1022–1029).

Fellbaum, C. (1998). *WordNet, an electronic lexical database*. Cambridge: MIT Press.

Giannakidou, E., Koutsonikola, V., Vakali, A., & Kompatsiaris, I. (2008). Co-clustering tags and social data sources. In *Proceedings of the 9th international conference on Web-age information management, China* (pp. 317–324).

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2007). Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing, 11*(6), 36–45.

Hotho, A., Jaschke, R., Schmitz, C., & Stumme, G. (2006a). Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web conference, LNCS* (Vol. 4011, pp. 411–426). Budva: Springer.

Hotho, A., Jaschke, R., Schmitz, C., & Stumme, G. (2006b). Trend detection in folksonomies. In *Proceedings of the 1st international conference on semantics and digital media technology* (Vol. 4306, pp. 56–70). Athens, Greece.

Kleinberg, J. (2006). Temporal dynamics of on-line information streams. In M. Garofalakis, J. Gehrke, & R. Rastogi (Eds.), *Data stream management: Processing high-speed data streams*. Springer.

Koutsonikola, V., Petridou, S., Vakali, A., Hacid, H., & Benatallah, B. (2008). Correlating time-related data sources with co-clustering. In *Proceedings of the 9th international conference on Web information systems engineering* (pp. 264–279). Auckland: Springer.

Koutsonikola, V., Vakali, A., Giannakidou, E., & Kompatsiaris, I. (2009). Clustering of social tagging system users: A topic and time based approach. In *10th int. conf. WISE* (Vol. 5802, pp 75–86). Berlin: Springer.

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations and applications. In J. Glaz & N. Balakrishnan (Eds.), *Recent advances on scan statistics and applications* (pp. 303–322).

Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 16–22). New York: ACM.

Nanopoulos, A., Gabriel, H., & Spiliopoulou, M. (2009). Spectral clustering in social-tagging systems. In *10th int. conf. on Web information systems engineering* (pp. 87–100).

Petridou, S. G., Koutsonikola, V. A., Vakali, A. I., & Papadimitriou, G. I. (2008). Time-aware web users' clustering. *IEEE Transactions on Knowledge and Data Engineering, 20*, 653–667.

Porter, M. F. (1997). An algorithm for suffix stripping. In *Readings in information retrieval* (pp. 313–316). San Francisco: Morgan Kaufmann Publishers Inc.

Rattenbury, T., Good, N., & Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 103–110). New York: ACM.

Richeldi, M., & Rossotto, M. (1995). Class-driven statistical discretization of continuous attributes (extended abstract). In *ECML'95* (pp. 335–338).

Russell, T. (2006). Cloudalicious: Folksonomy over time. In *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (pp. 364–364). Chapel Hill: ACM.

Shepitsen, A., Gemmell, J., Mobasher, B., & Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on recommender systems, RecSys '08* (pp. 259–266). Lausanne: ACM.

Sigurbjornsson, B, & van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web* (pp. 327–336). Beijing: ACM.

Specia, L., & Motta, E. (2007). Integrating folksonomies with the semantic web. In *4th ESWC* (pp. 624–639). Austria.

Sun, A., Zeng, D., Li, H., & Zheng, X. (2008). Discovering trends in collaborative tagging systems. In *Proceedings of the IEEE ISI 2008 PAISI, PACCF, and SOCO international workshops on intelligence and security informatics* (pp. 377–383). Berlin: Springer.

Swan, R., & Allan, J. (1999). Extracting significant time varying features from text. In *Proceedings of the eighth international conference on information and knowledge management* (pp. 38–45). New York: ACM.

Theodosiou, T., Angelis, L., Vakali, A., & Thomopoulos, G. (2007). Gene functional annotation by statistical analysis of biomedical articles. *International Journal of Medical Informatics, 76*(8), 601–613.

Vlachos, M., Meek, C., Vagena, Z., & Gunopulos, D. (2004). Identifying similarities, periodicities and bursts for online search queries. In *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on management of data* (pp. 131–142). New York: ACM.

Wetzker, R., Plumbaum, T., Korth, A., Bauckhage, C., Alpcan, T., & Metze, F. (2008a). Detecting trends in social bookmarking systems using a probabilistic generative model and smoothing. In *Proceedings of 19th international conference on pattern recognition (ICPR 2008)* (pp. 1–4). Piscataway: IEEE.

Wetzker, R., Zimmermann, C., & Bauckhage, C. (2008b) Analyzing social bookmarking systems: A del.icio.us cookbook. In *Proceedings of the ECAI 2008 mining social data workshop (2008)* (pp. 26–30).

Wu, E. H., Ng, M. K., Yip, A. M., & Chan, T. F. (2004). Discretization of multidimensional web data for informative dense regions discovery. In *Computational and information science* (pp. 718–724).

Wu, Z., & Palmer, M. (1994). Verm semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics* (pp. 133–138). New Mexico, USA.

Zhou, M., Bao, S., Wu, X., & Yu, Y. (2007). An unsupervised model for exploring hierarchical semantics from social annotations. In *Proceedings of the 6th international Semantic Web conference, (ISWC '07)* (pp. 680–693). Busan, Korea.