SEMSOC: Semantics Mining on Multimedia Social Data Sources

Eirini Giannakidou
Aristotle University of Thessaloniki
eirgiann@csd.auth.gr

Ioannis Kompatsiaris
Informatics and Telematics Institute, CERTH
ikom@iti.gr

Athena Vakali
Aristotle University of Thessaloniki
avakali@csd.auth.gr

Abstract

A huge amount of data and metadata emerges from Web 2.0 applications which have transformed the Web to a mass social interaction and collabo-ration medium. Collaborative Tagging Systems is a typical, popular and promising Web 2.0 application and despite its adoption it faces some serious limitations that restrict their usability. These limitations (no structure on tags, tags validation, spamming and redundancy) are more evident in the case of multi-media content due to its challenging automatic annotation and retrieval requirements. In this paper, we present an approach for social data clustering which combines jointly semantic, social and content-based information. We propose an unsupervised model for efficient and scalable mining on multimedia social-related data, which leads to the extraction of rich and trustworthy semantics and the improvement of retrieval in a social tagging system. Experimental results demonstrate the efficiency of the proposed approach.

1. Introduction

"Web 2.0" term is used to describe a group of technologies and web frameworks in which collaborative methods of information creation and organization are applied [1]. The key factor of its success is its constant update and continuous evolution, realized by users, who are treated as co-developers, since they provide data and metadata themselves dynamically in a continuous pace. As a result, the knowledge in these systems is built incrementally (by many users) in an evolutionary and decentralized manner, yielding in Emergent Semantics (as described the bibliography [2]).

A typical Web 2.0 application that has recently gained widespread popularity is the Collaborative Tagging Environments, where users label digital resources, by using freely chosen textual descriptions (tags). The simplicity and the user-centered design of those systems, have encouraged many web users to annotate their data by using tags which have proven to be very advantageous, especially, for search and retrieval in non-textual Web sources, such as photos, videos, etc. As a result, rapidly and in a short time, a huge amount of data and metadata became available in the Web. This user-driven approach of information creation and organization is known as folksonomy, a neologism proposed by Th. Vander Wal in [3], and its real strength lies in the fact that its structure and dynamics are similar to those of a complex system, yielding in stable and knowledge-rich patterns after a specific usage period ([4], [5]).

While social data (i.e. folksonomies) seem very promising sources of information, they have some serious limitations that restrict their usability [6]. First of all, users are prone to make mistakes and they often suggest invalid metadata (*tag spamming*). Additionally, the lack of (hierarchical) structure of information results in *tag ambiguity* (a tag may have many senses), *tag synonymy* (two different tags may have the same meaning) and *granularity variation* (users do not use the same description level, when they refer to a concept), which restricts the retrieval ability of such systems. People tend to use redundant tags, in order to tackle low recall, but this worsens the precision of the system, as it causes many irrelevant objects to be fetched to the users.

A current research trend to tackle the above mentioned limitations and extract semantics from social data is to employ clustering, as an unsupervised model, that separates the resources into meaningful groups (i.e. clusters). Each cluster corresponds to a particular topic-domain and the set of its containing tags reflects the way users perceive a particular domain.. In this context, here, we propose an approach which jointly considers social, semantic and content features to cluster multimedia data sources. More specifically, to cluster multimedia social sources, we combine knowledge (tag co-occurrences, user interactions, tag assignments, etc.) derived from a folksonomy (i.e. *social knowl-*

edge) with already defined and widely used ontologies (i.e. *semantic knowledge*) and content-based information (visual or audio low-level features) extracted by content analysis techniques (i.e. *content-related information*). Thus, we put emphasis on using all available "tracks" of knowledge (namely the semantic, the social and the content) to perform clustering. This approach is followed in an effort to result in better, more meaningful and "pure" clusters (as will be shown in Section 7). By having such improved clusters, we result in more trustworthy emergent semantics and the overall performance of knowledge extraction is improved and facilitated.

The structure of the paper is as follows: Related work is presented in Section 2. Section 3 gives an introduction to the basics of social tagging systems, while Section 4 specifies the motivation for our work. In Section 5 the formulation of the problem addressed in this paper based on joint semantic, social and content knowledge is described. The implementation of the specific approach, called SEMSOC, and its modules follow in Section 6. Finally, experimental results and conclusions are presented in Section 7 and 8, respectively.

2. Related Work

Many earlier research efforts have focused on exploiting knowledge stored and often "hidden" in folksonomies and they have dealt with the following topics:

- *clustering techniques* based only on tagging information and tag co-occurrence to derive semantically-related groups of tags and resources, out of a folksonomy, which are met in [8], [9], [10] and in Flickr[1] clusters. Such methodologies involve only tag statistical analysis and they lack of any semantic information that could guide the clustering process. Thus, they quite often yield clusters of co-occuring tags, which cannot be mapped to an actual topic and cannot be interpreted by a user. Additionally, they don't always tackle quite well the tag synonymy issue, since synonymous tags are commonly given by different users and they seldom co-occur.
- *ontology-driven* tagging organization and mining, by combining Web 2.0 and Semantic Web [11] ideas. Such efforts include building of an ontology that formalizes the activity of tagging, so as to enable the exchange, comparison and reasoning over the tag data acquired from varied tagging applications [12], and a number of approaches which have focused mainly on the exploration of the tag space and the detection of emergent relations in social data, which will be exploited for ontology building and/or evolution ([13], [14], [15], [16], [17], [18]). Clustering based on tag co-occurrence is utilized in the latter approaches.
- *content-based* analysis on tagging-related sources, such as in [19] where a method is introduced for exploiting both tags and visual features (in a supplementary manner) for browsing and retrieving of semantically related images. The authors claim that content-based analysis is able to tackle the intrinsic shortcomings of a multimedia collaborative tagging system and can contribute to the emergence of interesting (semantic) relationships between data sources. Other efforts to design tools that employ simple image analysis algorithms and apply them on Flickr images have appeared in [20], [21], yet still they work separately and are not intended for semantic similarity extraction or integrated navigation in the social tagging system.

Despite the active research efforts in this area, the full potential of Web 2.0 data management has not been exploited, yet. The aforementioned overview of existing approaches indicates that clustering is, quite often, employed as a first step to semantics mining of a folksonomy. Indeed, each cluster encompasses the collective users' view around a specific topic, which can be further exploited for semantics extraction. According to the authors' knowledge and as discussed above, all researchers rely solely on tagging data, in order to analyze and cluster folksonomies, ignoring the semantic aspects of tags. Similarly, few approaches exploit visual features during the clustering procedure; they are mostly used in browsing and retrieval applications. We claim that extracting the most representative numerical descriptors and defining appropriate similarity metrics that emulate the "human notion" of similarity can further contribute to more efficient social tag clustering. In the sequel, we describe our method, which fuses tag co-occurrence with semantic knowledge and low-level descriptors of the resources, so as to get fine clusters of multimedia social data.

3. Social Tagging Systems Basics

A Social Tagging System, STS, is a web-based application, where users assign tags (i.e. arbitrary textual descriptions) to digital resources. The digital resources are either uploaded by users or, are, already, available in the web. The users are either "isolated" or, more commonly, members of web communities (i.e. social networks) and their main motivation (for tagging) is information organization and sharing. The tagging activity inside an STS shows the way users categorize

[1] http://www.flickr.com

resources and it is known as its folksonomy [3]. Figure 1 depicts the basic structure of a web-based STS.

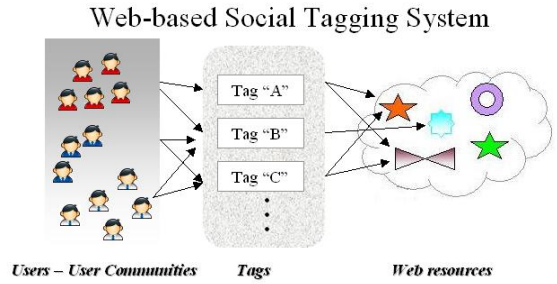

Figure 1. A web-based social tagging system

The most common definition of a social tagging system has been quoted in [15] and we adopt it here as well:

Definition 1: Given an STS, its derived folksonomy, **F**, is defined as the tuple: $\mathbf{F} = (U, R, T, A)$, where U, R, T, A are finite sets describing the groups of users, resources, tags and user annotations (i.e. tag assignments) in a STS, respectively. Specifically, the users' annotations set A is modeled as a triadic relation between the other sets (i.e. $A \subseteq U \times R \times T$).

Each STS handles a particular kind of resources. For instance, Flickr handles photos, while del.icio.us (found at: http://del.icio.us) handles urls, YouTube (found at: http://www.youtube.com) handles videos, etc. Nevertheless, resource management by a STS is a transparent process, which does not rely on the varying nature of digital resources (i.e. text or multimedia). Specifically, each resource R_j in a STS is associated only with user-generated metadata (produced through the tagging activity), regardless of the specific nature of R_j. These involve: i) some context information, such as the user who uploaded the specified resource, the users who annotated it, the time when each of the above tasks occurred etc., and ii) the group of tags assigned to it. Even in the case of multimedia resources, no analysis techniques for intrinsic feature extraction are employed by STS, in alignment with one of the main principles of Web 2.0, which calls for simplicity and use of lightweight structures.

In this paper, we focus on tag metadata and leave context metadata contribution for future work. We consider that the context of each resource is captured by the manifold annotations it has received; hence, we characterize and define resources by their corresponding tags:

Definition 2: Each resource R_j in a STS is represented by aggregating the tags assigned to it by all users and it is identified by:

$$R_j = (a_1 * \text{tag}_{j1}, a_2 * \text{tag}_{j2}, \ldots, a_z * \text{tag}_{jz}) \quad , \quad (1)$$

where z is the number of tags assigned to the resource by all users and the coefficients α_i denote the number of times the tag_{ji} has been used in R_j's annotation.

4. Motivation for Multimedia Retrieval

STS have played a crucial role in the improvement of handling and utilization of multimedia resources. In fact, this was a key factor for their wide spread and adoption by the web community, since the retrieval of such resources has long been extremely difficult, without proper metadata. Employing experts to perform annotations is an expensive and practically immutable procedure. On the same time, despite the recent progress in content-based automatic extraction of semantic metadata from multimedia, such techniques are far from being perfect and generic applicable [22].

This can be overcome by exploiting the annotations (tags) given in a STS and hence receiving readily and without cost user generated metadata that best fits the community point of view of the specific resources. In this way, handling of multimedia data becomes a tag-oriented procedure and the extraction of their context (i.e. semantics) for their analysis turns into the problem of extracting the semantics and analyzing of their corresponding tags. In the following section a joint approach that utilizes both the social and semantic aspects of the tags of multimedia resources as well as the content of the resources is presented.

5. Problem Formulation

In our work, we introduce a *two-step* approach for clustering on multimedia resources over a STS. In the *first step* tags guide our proposed clustering process. Since we characterize multimedia resources as sets of tags, the semantic and the social aspects of tags will be taken into consideration and analyzed. In the *second step*, we employ content-based analysis of the resources, in order to minimize as possible the undesirable effects of tag invalidity and tag ambiguity and refine the extracted clusters. The intrinsic features (content-related) of the resources in every cluster are examined and the distant objects are removed from the cluster. In sections 5.1 and 5.2 a problem formulation is quoted, to emphasize the required concept definitions and the mathematical notations used in the first and second step of the process, respectively.

5.1. Joint Semantic & Social Data Clustering

Given, a set of resources R where $|R| = N$, clustering organizes the N resources into k clusters C_1, C_2,

…, C_k, with respect to an Attribute Set, *AS*, in which each element is a so called attribute (feature or dimension) and it is used to measure the similarity between the resources. Resources assigned in a cluster should be strongly similar to each other, according to some metric of similarity, while the ones assigned to different clusters should be dissimilar [23].

In a STS, where the resources are expressed via the tags assigned to them (see Eq. 1), we adopt a clustering approach in which a set of representative, distinguishing tags will form the so called attribute set, *AS*, such that two resources with tags of strong similarity will be grouped together. We define an association function *Similarity Factor* sf_{ji} between a resource R_j and an attribute $attr_i$, which is evaluated by encompassing both social and semantic relations between the resources' tags and the tag that corresponds to the specified attribute. This joint proposed approach aims at producing homogeneous resources clusters.

As introduced in Section 2, current STS, which employ clustering, rely solely on tag co-occurrence, to estimate tag closeness, and, hence, resource closeness. We refer to such a similarity between two tags as *social similarity, SoS* and we define it as follows:

$$SoS(t_x,t_y)=\frac{\sum_{r_j\in R}(u_z,t_x,r_j)\cap(u_w,t_y,r_j)}{max\left(\sum_{r_j\in R}(u_z,t_x,r_j),\sum_{r_j\in R}(u_w,t_y,r_j)\right)} \quad (2)$$

$\forall(u_w,t_x,r_j),(u_z,t_y,r_j)$, where $u_w,u_z\in U$.

In order to estimate semantic similarity between tags, external resources i.e. semantic web ontologies, thesauri and/or lexicons available in the web need to be employed. A mapping technique is applied to act as a bridge between a free-text tag and a structured concept of the used resource. There are a number of available measures that attempt to evaluate the semantic distance between ontology concepts, and a very thorough presentation of the most widely used ones can be found in [24]. We adopted the Wu & Palmer measure, due to its simple and straightforward application on our data. Based on this, the semantic distance between two concepts is proportional to the path distance between them. Thus, let t_x and t_y be two tags for which we want to find the semantic similarity and $\vec{t}_x,\vec{t}_y$ be their corresponding mapping concepts via the ontology *O*. Then, their Semantic Similarity, SeS, is calculated as:

$$SeS(t_x,t_y)=\frac{2\times depth(LCS)}{\left[depth(\vec{t}_x)+depth(\vec{t}_y)\right]} \quad (3)$$

where $depth(\vec{t}_x)$ is the maximum path length from the root to $\vec{t}_x$ and *LCS* is the least common subsumer of $\vec{t}_x$ and $\vec{t}_y$.

Hence, we can estimate the overall similarity between two tags, which constitutes the combination of their social and semantic similarity. In order to examine the impact that each kind of knowledge (social or semantic) has on the resources clustering, we join them in the form of a weighted sum. Specifically, a factor w is employed to define the effect each track has on the estimation of their joint similarity. Thus, we define the *Similarity Score*, *SS* between t_x and t_y in terms of both their social (Equation 2) and semantic (Equation 3) similarity as:

$$SS(t_x,t_y)=w*SoS(t_x,t_y)+(1-w)*SeS(t_x,t_y) \quad (4)$$

where $w\in[0,1]$ and is a normalization parameter which adjusts the magnitude of the semantic similarity against the social one upon the final outcome.

Having specified the similarity metric between tags, we can proceed to the estimation of similarity factors, sf_{ji}, discussed in the beginning of the Section.

Definition 3 "Similarity Factor": Given a resource R_j, in which the users have assigned $|R_j|$ tags, and an attribute $attr_i$, we define a Similarity Factor, sf_{ji}, between the specified resource and the specified attribute, the maximum *SS* between every tag assigned to resource R_j and the attribute $attr_i$. Thus:

$$sf(R_j,attr_i)=max_{x=1..|R_j|}\{SS(t_x,attr_i)\}, \quad (5)$$

where $R_j\in R,t_x\in T,attr_i\in AS$.

In the above definition, we assume that all the tags assigned to each resource are relevant to the content. Alternatively, taking the average SimScore could be more robust against tag-spamming, but it would be biased against resources which receive tags of different kinds (i.e. regarding a "sea" attribute, a resource with a tag "sea" would get higher score than another resource with tags "sea", "beach", "anna", "2007", although both of them involve sea). In the 2nd step of the process (that content analysis is employed and described in the sequel), we take control of the tag-spamming issue and track the noisy tags that surpassed the first step, cleaning, thus, the clusters from resources with erroneous annotations.

The values of Similarity Factors between each of the *N* resources and *d* attributes are then used to form the $N\times d$ so-called Similarity Matrix, as follows:

$$SimMatrix(j,i)=sf(R_j,attr_i) \quad (6),$$

where $1\le j\le N,1\le i\le d$. This resources' similarity matrix is the input to the clustering procedure, out of which *k* resources clusters shall arise.

5.2. Data Refinement with Visual Features

To minimize the intrinsic shortcomings of STS, our technique is augmented by content analysis of the multimedia resources so another measure of similarity between resources is introduced, namely their Content Similarity. In order to estimate the content-based similarity, appropriate similarity metrics between numerical automatically extracted low-level features are used. Such features can be extracted from multimedia sources, using the MPEG-7 standard [25] which defines appropriate descriptors together with their extraction techniques and similarity matching distances. More specifically, the MPEG-7 eXperimentation Model, XM provides a reference implementation which can be used in our approach [26]. Here we proceed to SEMSOC second step which is based on identifying low-level features of the multimedia resources, which are extracted from images and form an image feature vector. The image feature vector proposed in this work involves two different descriptors of the MPEG-7 standard, namely the *Color Structure Histogram (CSH)* and *Edge Histogram (EH)* descriptors, chosen due to their effectiveness in similarity retrieval. Their extraction is performed according to the guidelines provided by the MPEG-7 XM and then, an *image feature vector* is produced, for every resource, by encompassing the extracted MPEG-7 descriptors in a single vector. Thus, the *Content Similarity* between two resources is the similarity of their corresponding image feature vectors.The distance functions used to calculate the content similarity are according to the guidelines of MPEG-7 and they are provided by the MPEG-7 XM. Based on content similarity, an outlier analysis is performed in every cluster, aiming at removing the most distant objects (which surpassed Step 1, mostly due to noisy tags). By this way, we will show that we result in more homogeneous clusters.

Out of each final extracted cluster a *tag cluster* and a *cluster topic* are extracted, as follows.

Definition 4 "Tag Cluster": Given a resource cluster *C*, we call *Tag Cluster*, *TC* the set with the user-assigned tags that describe the resources in *C*.

Definition 5 "Cluster Topic": Given a resource cluster *C*, we define its cluster topic as the tags that belong to its corresponding *Tag Cluster,* having frequency above a user-defined threshold τ.

6. SEMSOC: Implementing Semantic / Social / Content Clustering

We have implemented a framework, the SEMSOC framework (SEMantic, SOcial, Content similarity) that hosts the proposed multimedia resource clustering process. As described in the previous section, the clustering in SEMSOC is driven by a fusion mechanism which combines the semantic, social and content similarity of the resources. As depicted in Figure 2 the proposed framework realizes two steps: in the first step a tag-guided clustering of the resources takes place, based on semantic and social aspects of their accompanying tags and in the second step image analysis techniques are applied, in an effort to tackle the misleading-tag effect (tag spamming) and ameliorate the extracted clusters.

The following tasks are carried out under SEMSOC:

Data collection: The acquisition and storage of the multimedia content and its accompanying metadata (i.e. tags) from a STS are held by a crawler. The crawler starts with an initial (random) tag and collects multimedia resources that contain the specified tag in their annotation. After a specific number of downloads the initial tag is replaced by another one picked from last resource's annotation. Gathered metadata are stored in XML format. Having collected the resources with their accompanying metadata, the first step of the process follows. It comprises preprocessing, attribute selection, similarities evaluation and social/semantic clustering.

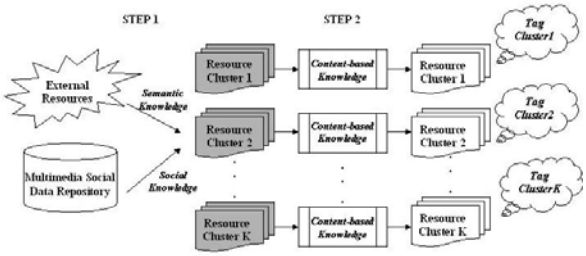

Figure 2. SEMSOC two-step framework

Step 1 - Preprocessing: The first step starts with preprocessing and cleaning tasks for the collected metadata. The raw tags are forwarded to a preprocessing module to get back normalized tags. Currently, SEMSOC supports a quite simple preprocessing of the gathered tags. Two tasks are involved in this process: a spelling normalization, so that different written forms of the same tag are mapped to the same normalized tag (e.g. Sea, sea) and a filtering such that the infrequent tags are filtered out as trivial.

Attribute selection and similarities evaluation: At the attribute selection process the top D representative processed tags are selected to form the attribute set.

Since the resource clustering procedure, in the first step, is tag-guided, it is reasonable that a set of distinguishing tags will form the attributes, based on which clustering will be performed. Since the number of distinguishing tags in the collected metadata is of a very large-scale, an attribute selection process must by defined. We employ the D most frequent tags (after the preprocessing phase) to form the attribute set. Then, the tag similarity scores calculation occurs. It comprises two sub-processes, namely the *Semantic Tag Similarity* and the *Social Tag Similarity*, and a fusion mechanism. The module takes as input the multimedia resources, as sets of tags, and the attribute set and outputs *similarity factors* between each resource and each attribute. The extracted similarity factors construct the *similarity matrix* upon which the application of clustering algorithms will take place at the next phase. More specifically, the module executes a cycle of operations per digital resource such that in each cycle the similarity factors of the specified resource are evaluated (see Section 5).

Social/Semantic Clustering: The first step of SEMSOC finalizes with the clustering procedure. The similarity matrix is used as input to the actual clustering process. It must be noted that SEMSOC framework adapts an open architecture which allows the application of any clustering algorithm.

Step 2 - Cluster refinement with Content-based analysis: The second step of the process involves refinement of the extracted clusters of the first step. Image analysis techniques are employed that aim at estimating visual (content) similarities between resources in the same cluster and tracking possible noisy tags. The visual heterogeneous resources of each (first-step) cluster (outliers) are removed and a new cleaner clustering yields.

7. Experimentation

In this section, experimental results of the application of the proposed SEMSOC approach to a corpus of multimedia resources obtained from an STS are presented.

A basic preparatory step of the process involves the acquisition of external resources that will be used for the mapping of tags to concepts and the evaluation of their semantic similarity. We used WordNet, since it is a worldwide used lexicon, which provides mechanisms for calculation of semantic distances between concepts [27].

To carry out the experimentation phase and the evaluation of our system, a dataset from Flickr was crawled. It consists of 3000 images (size 500x735) and includes images that depict *cityscape*, *seaside*, *mountain*, *roadside*, *landscape* and *sport-side* locations.

7.1. Clustering results

To ensure the stability and robustness of clustering results, a variety of clustering algorithms were tested. Specifically, we used a partitional algorithm (K-means), a hierarchical (Agglomerative) and a conceptual clustering process (Cobweb) [23]. To evaluate the quality of the extracted clusters of resources, for each technique described in the paper, each image resource was manually annotated with respect to the emergent cluster topics (see Section 5). Then, we use precision (Pr) and recall (R) as follows. Let, C be an extracted cluster and t its emergent CT. We call RR the set of corpus resources that have received manual annotations that match t (i.e. Relevant Resources). We define i) precision as the fraction of resources that belong to C and are relevant (i.e. $Pr = |C \cap RR| / |C|$), and ii) recall as the fraction of relevant resources which belong to C (i.e. $R = |RR \cap C| / |R|$). In Tables 1 and 2 the precision (Pr) and recall (R) of the clustering algorithms are quoted for different values of number of clusters, respectively. In each table, the measure is calculated at each step of SEMSOC separately. It can be seen that K-means and Hierarchical had both satisfying performance, while Cobweb was worse. Furthermore, the outcome shows clearly that content-related knowledge (employed in step 2) improves the quality of the extracted clusters, without deteriorating the recall of the system.

Table 1. Precision in each step of SEMSOC for varying algorithms and varying number of clusters (K)

Algorithms	K = 14		K = 17		K = 20	
K-means	0.657	0.77	0.75	0.813	0.687	0.806
Hierarchical	0.679	0.842	0.744	0.85	0.675	0.752
Cobweb	0.552	0.723	0.65	0.708	0.589	0.673
	Step 1	*Step 2*	*Step 1*	*Step 2*	*Step 1*	*Step 2*

Table 2. Recall in each step of SEMSOC for varying algorithms and varying number of clusters (K)

Algorithms	K = 14		K = 17		K = 20	
K-means	0.6	0.57	0.781	0.75	0.634	0.6
Hierarchical	0.71	0.69	0.566	0.566	0.694	0.652
Cobweb	0.749	0.739	0.805	0.78	0.78	0.732
	Step 1	*Step 2*	*Step 1*	*Step 2*	*Step 1*	*Step 2*

Furthermore, different values for the number of attributes were tested. The results, calculated after step 2, are in Table 3 and show that, by increasing the number of attributes, a better clustering yields. Testing of other approaches for attribute selection is in our

approaches for attribute selection is in our plans for future work and expected to improve more the quality of clustering.

Table 3. Precision and Recall of varying clustering algorithms, using different number of attributes (D)

Algorithms	D = 30		D = 60	
K-means	0.813	0.75	0.85	0.748
Hierarchical	0.85	0.69	0.89	0.73
Cobweb	0.723	0.78	0.81	0.68
	Pr	*R*	*Pr*	*R*

Generally, most of the clusters the system generated were homogeneous and meaningful. The corresponding tag clusters were also very representative and highly informative. In Figure 3 two indicative snapshots of clusters and tag clusters are shown. The clusters shown are the outcome of Step 1 and the resources surrounded by a red box are removed during the Step 2 of the process. It must be noted that for space reasons, only hierarchical algorithm snapshots are shown.

Figure 3. Snapshot of a sea and a tennis cluster with their emergent TC - Identification of a misleading tag in the sea cluster and rejection of the resource (surrounded by a red box)

Figure 4. Snapshot of a France cluster, TC = {france, paris, louvre…}

Surprisingly, some unexpected clusters revealed, whose topics neither are included in the domains that form the dataset nor can be derived by the latter (i.e. related domain). An indicative cluster of this kind is shown in Figure 4 and depicts a cluster of images about France. This reflects the strong associations between images that users have enforced through their tags.

7.2. Scenarios

In this section we will show that the proposed SEMSOC approach tackles quite well the shortcomings of an STS, described in the Introduction. Furthermore, its ability in subdomain identification within a domain is demonstrated. (Due to space restriction only some snapshots are shown indicatevely).

Tag ambiguity: The clustering algorithms handled well the specified issue and distinguished different senses of the same tag, by dividing the corresponding resources into different clusters (see Figure 5).

Questionable reliability: It is expected that misleading tags in some annotations are practically overwhelmed by the massive activity of a large number of users. Nevertheless, the content similarity factor, employed in step 2 of the process, tracks and removes from a cluster those objects that have a visual appearance very different from the rest ones (e.g. Figure 3).

Fig. 5. Different clusters for the ambiguous tag: *wave* (a) members of cluster with TC = {wave, sea, water,…} (b) members of cluster with TC = {wave, signal, hand, person, …}

Tag redundancy and lack of hierarchical relations: Since semantic similarity of tags is employed, tag redundancy is no more needed. The system inherits the structure of the external resource used (i.e. the structure of concepts of WordNet).

Identification of subdomains: SEMSOC accomplishes to find meaningful sub-clusters, inside a generic cluster. For instance, the initial group of Roadside images is split by the process into three more specific clusters, depicted in Fig. 6., with (a) TC = {building, roof, street,…}, (b) TC = {car, race, Porsche, street,…} (c) TC = {caribbean, carnival, festival, people, street,…}.

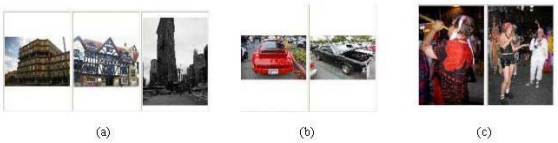

Figure 6. Members of different clusters of Roadside

8. Conclusions

In this paper, a robust clustering method for social data grouping which relies jointly on social, semantic and content knowledge was presented. This is different

from the usual case where only tagging data are used, in order to analyze and cluster folksonomies. Our method has a number of potential applications. Apart from the obvious retrieval applications, the tag clusters produced can be used for semantics extraction and knowledge mining, in general and more specifically in automated multimedia content analysis, being used for example as training sets for specific concepts represented by tags. Future work includes the incorporation of visual features in the clustering procedure, based on using a common input vector resulting from all the available information per resource. In order to achieve this, appropriate normalization techniques need to be employed.

9. References

[1] O'Reilly, T. "What is Web 2.0," Published on http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (September 30, 2005)

[2] Steels, L., Semiotic Dynamics for Embodied Agents. *IEEE Intelligent Systems*, 21 (3): 32-38 (2006).

[3] Vander Wal, Th., (2005, Feb.) Explaining and showing broad and narrow folksonomies. Blog post 2005-02-21. http://www.vanderwal.net/random/category.php?cat=153

[4] Cattuto, C., Collaborative tagging as a complex system. Talk given at *International School on Semiotic Dynamics*, Language and Complexity, Erice (2005).

[5] Halpin, H. and Shepard, H., Evolving Ontologies from Folksonomies: Tagging as a Complex System. http://www.ibiblio.org/hhalpin/homepage/notes/taggingcss.html (visited Oct 2007).

[6] Golder, S. and Huberman, A., The Structure of Collaborative Tagging Systems. *Journal of Information Science*, (2006).

[8] Begelman, G., Keller, Ph., Smadja, F., Automated Tag Clustering: Improving search and exploration in the tag space. *Collaborative Web Tagging Workshop*, (2006)

[9] Grahl M., Hotho A. and G. Stumme. Conceptual Clustering of Social Bookmarking Sites. *7th International Conference on Knowledge Management*, 356-364, Know-Center,Graz, Austria, 2007.

[10] Jaschke R., Hotho A., Schmitz Ch., Ganter B. and G. Stumme. TRIAS - An Algorithm for Mining Iceberg Tri-Lattices. *Proc of the 6th IEEE International Conference on Data Mining*, 907-911, 2006.

[11] Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American*, 34-43 (2001).

[12] Gruber, T., Folksonomy of Ontology: A Mash-up of Apples and Oranges. First On-Line conference on Metadata and Semantics Research MTSR (2005).

[13] Schmitz, P., Inducing Ontology from Flickr Tags. In *Proc. of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, (2006).

[14] Mika, P., Ontologies are Us: A Unified Model of Social Networks and Semantics. In *Proc. of the 4th International Semantic Web Conference*, (2005).

[15] Schmitz, C., Hotho, A., Jaschke, R., Stumme, G.. Mining Association Rules in Folksonomies. In *Proc. of the (IFCS 2006)*, pages 261-270, Ljubljana, (2006).

[16] Specia, L., and Motta, E., Integrating Folksonomies with the Semantic Web. In *Proc. of the 4th European Semantic Web Conference*, (2007).

[17] Wu, X., Zhang, L., Yu, Y., Exploring Social Annotations for the Semantic Web. In *Proc. of the 15th WWW Conference*, Edinnburgh, Scotland, (2006).

[18] Zhou, M., Bao, S., Wu, X., and Yu Y., An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations. In *Proc. of the 6th International Semantic Web Conference*, (2007).

[19] Aurnhammer, M., Hanappe, P., Steels, L., Augmenting navigation for collaborative tagging with emergent semantics. In *Proc. of the 5th ISWC* (2006)

[20] Bumgardner, J.: Experimental colr pickr. http://www.krazydad.com/colrpickr/ (2006)

[21] Langreiter, C.: Retrievr. http://labs.systemone.at/retrievr/ (2006)

[22] Hobson, P., Kompatsiaris, Y. Advances in semantic multimedia analysis for personalised content access, *IEEE International Symposium on Circuits and Systems* (2006).

[23] Xu, R., Survey of Clustering Algorithms. In *IEEE Transactions on Neural Networks*, Vol.16, No.3, May 2005.

[24] Maguitman, A., Lord, P.W., Menczer, F., Roinestad, H., Vespignani, A., Algorithmic Detection of Semantic Similarity. In *Proc. of 14th WWW Conference* (2005).

[25] ISO/MPEG N3752 - Martínez, J.M. "Overview of the MPEG-7 Standard (v4.0)"

[26] MPEG-7 Visual Experimentation Model (XM), Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4062, Mar., 2001.

[27] Fellbaum, C.. WordNet, an electronic lexical database. The MIT Press (1990).