Harvesting Intelligence in Multimedia Social Tagging Systems

Eirini Giannakidou, Foteini Kaklidou, Elisavet Chatzilari, Ioannis Kompatsiaris, and Athena Vakali

Abstract As more people adopt tagging practices, social tagging systems tend to form rich knowledge repositories that enable the extraction of patterns reflecting the way content semantics is perceived by the web users. This is of particular importance, especially in the case of multimedia content, since the availability of such content in the web is very high and its efficient retrieval using textual annotations or content-based automatically extracted metadata still remains a challenge. It is argued that complementing multimedia analysis techniques with knowledge drawn from web social annotations may facilitate multimedia content management. This chapter focuses on analyzing tagging patterns and combining them with content feature extraction methods, generating, thus, intelligence from multimedia social tagging systems. Emphasis is placed on using all available "tracks" of knowledge, that is *tag co-occurrence* together with *semantic relations* among tags and *low-level features* of the content. Towards this direction, a survey on the theoretical background and the adopted practices for analysis of multimedia social content are presented. A case study from Flickr illustrates the efficiency of the proposed approach.

Eirini Giannakidou
Informatics Department, Aristotle University of Thessaloniki, Greece, e-mail: eirgiann@csd.auth.gr
Informatics & Telematics Institute, CERTH, Thermi - Thessaloniki, Greece e-mail: igiannak@iti.gr

Foteini Kaklidou
Informatics & Telematics Institute, CERTH, Thermi - Thessaloniki, Greece e-mail: fkaklid@iti.gr

Elisavet Chatzilari
Informatics & Telematics Institute, CERTH, Thermi - Thessaloniki, Greece e-mail: echatzi@iti.gr

Ioannis Kompatsiaris
Informatics & Telematics Institute, CERTH, Thermi - Thessaloniki, Greece e-mail: ikom@iti.gr

Athena Vakali
Informatics Department, Aristotle University of Thessaloniki, Greece, e-mail: avakali@csd.auth.gr

1 Introduction

Participating users in the web act as co-developers and their actions and interactions with one another have produced a valuable, quite difficult to handle, though, information repository, enhanced with social characteristics, derived from the communication and the collaboration among them. This social dimension was emphasized in the next generation of web, namely Web 2.0 or Social Web technologies and applications [1], resulting in a remarkable bursting of web usage and content availability, and addressing, at the same time, the need for efficient techniques' deployment for exploiting this collective knowledge.

Central to this new web is the concept of tagging (i.e. users attaching keywords to describe digital data sources). The process of having end-users adding their own metadata to internet resources, namely *social* or *collaborative tagging*, introduces a new way of digital data sources' organization and retrieval in the web that constitutes the core process in a number of web 2.0 applications that have received tremendous attention lately, such as Flickr[1], del.icio.us[2], YouTube[3], Technorati[4] and so on. The remarkable with tagging activity is that although completely subjective and without relying on a controlled vocabulary, it has dynamics similar to those of a *complex system* [2], [3], i.e. knowledge is built incrementally in an evolutionary and decentralized manner, yielding stable and knowledge-rich patterns, namely Emergent Semantics [4]. Thus, unlike earlier static knowledge representation structures, social tagging systems are dynamic and have a noteworthy ability in capturing the community' s point of view of the specific data sources and the general trends, at a given time. Additionally, they capture social relations between the community members. Therefore, they constitute promising data structures for knowledge mining.

In this chapter, we study social tagging systems that host multimedia data sources. We argue that the metadata given by users in social tagging systems (i.e. tags) form a valuable knowledge source which has a social dimension and is extremely dynamic, since users add content and tags all the time. Towards this direction, we focus on analyzing tagging patterns and combining them with content feature extraction methods, in order to get useful knowledge about the content, that will facilitate its retrieval. This knowledge can be regarded as a first basic step towards intelligence generation from multimedia social tagging systems. The problem to be analyzed in this chapter is how to exploit this source and overcome, at the same time, the intrinsic limitations these systems have and are summarized in *i) tag redundancies and ambiguities*, raised by the complete lack of structure and hierarchical relations, and *ii) metadata questionable validity*, as users are prone to make mistakes.

[1] Flickr photo-sharing system: *http://www.flickr.com*

[2] Del.icio.us social bookmarks manager: *http://del.icio.us*

[3] YouTube video-sharing website: *http://www.youtube.com*

[4] Technorati blog search engine: *http://technorati.com*

Our methods are based on developing solutions for linking descriptive semantics, yielded by tag processing, with the low-level features of the media assets. In order to derive such semantics from tags and get information interpretable by the end user, a clustering procedure takes place. Clustering is often employed in the bibliography of social tagging systems as a way of grouping together tags related to a certain topic. Here, we put emphasis on using all available "tracks" of knowledge, namely *social knowledge* (i.e. knowledge that can be derived from tagging systems, e.g. tag co-occurrence), *semantic knowledge* (i.e. knowledge about the meaning of the concepts e.g. hierarchical relations among them), and *content-based knowledge* (i.e. the low-level features of the multimedia data). Our goal is to yield useful knowledge from the multitude of user annotations, which, especially in the case of multimedia data, can be used to semantically enrich the specified content and facilitate the retrieval task, promoting, thus, its exploitation. A case study on 10000 and 3000 resources from Flickr is used to demonstrate that the exploitation of users' annotations produces semantic metadata and provides added-value to the available multimedia content.

The structure of this chapter is as follows. Section 2 gives a short overview to the multimedia content annotation approaches and introduces multimedia social tagging systems, emphasizing on the reasons of their popularity. An extended state-of-the-art follows, in Section 3, including *i)* approaches that analyze and/or cluster social tagging systems, *ii)* content-based multimedia techniques and, *iii)* cases in which the two methods are combined. In Section 4 our approach, which joins tagging and content-based knowledge, is presented. Next, experimental results and use cases of the proposed approach are quoted in Section 5 and 7, respectively. Finally, Section 7 concludes the chapter.

2 Multimedia Content Annotation

Multimedia is, increasingly, gaining popularity in the web with several technologies supporting the use of images, animation, video and audio to supplement the traditional medium of text. The basic reason behind the vast quantity of multimedia web data was the rapid technological growth, together with some quality traits that the combined use of multiple modalities gives to the content, such as natural design, interactivity and pleasure to work with. In order for that enhanced-valued content to be easily found and accessible, special design/management discipline is required. Unconstrained use of multimedia results in a chaotic web environment that confuses users and makes it hard for them to locate the information they are interested in.

There is a growing number of research methods for analyzing, understanding and delivering multimedia content which are based on content-based features extracted from the multimedia data. These methods rely on extracting low-level features of the digital objects either for retrieval by visual similarity or for associating them with high-level concepts. While automatic extraction of low-level features and mapping to high-level concepts is possible in many applications, their major drawback, lies in the distance between the high-level concepts that describe the multimedia content

and the extracted low-level features, a problem that is known as the *semantic gap* [5]. Semantic gap is a serious concern in these methods, as it makes retrieval by semantic relevance a very difficult task. Therefore challenging methods for efficient mapping to a large number of high-level concepts are needed.

Another approach to multimedia content handling is based on utilizing additional knowledge about the content, given in the form of *metadata*. Metadata is defined in [6] as *"structured information that describes, explains, locates, or otherwise helps in retrieving, using or managing a resource"*. Thus, retrieval of multimedia content may be based on its metadata, exclusively, or on complementing existing content-based approaches with accompanying content metadata. However adding metadata to content still remains an expensive and difficult to maintain and evolve process, as it requires a group of experts spending human-hours in manually annotating the content. Moreover, the defined metadata reflect the experts' point of view of the particular content, which is not always identical with the users' perception of it. With the enormous growth of multimedia content and the rapid changes in the web environment, a more dynamic approach is needed that ensures that the metadata provided encompass the user community' s awareness and understanding of the available content. We argue that such metadata can be drawn from multimedia social tagging systems, which are web-based applications that allow users upload/share/browse multimedia content and annotate it by completely freely chosen metadata. A more detailed description of multimedia social tagging systems follows.

2.1 Multimedia Social Data Sources

Given the warm embrace of tagging activity by web users,currently a variety of social tagging systems prevail in the web map. These systems are web-based applications in which users add textual descriptions (i.e. tags) to digital content (i.e. resources), enriching it, thus, with ready-to-use metadata and making its retrieval more efficient. Users may participate as atoms or, more commonly, as members of communities. The resources of these systems are specified/uploaded by users and may be available to the entire web community, along with their metadata. There is no restriction on the selection of tags; any user may choose any term that is meaningful to him/her and thinks as appropriate for the resource description. This rough description illustrates the 3-partite structure of social tagging systems, which is depicted in Figure 1.

Adding keywords (i.e. tags) to data sources is not something new. Librarians and indexers have been using keywords to facilitate the retrieval of their resources, a long time ago. Ever since many professionals have adopted the tagging technique in an effort to organize and enhance searching in their data [7]. The feature that is new in social tagging systems and promoted their endorsement by the majority of web community is that tagging is now performed by everyone, not only by a small group of experts, and that the tags are being made public and shared to anyone. This

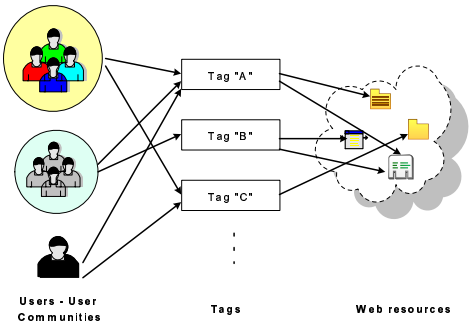

Fig. 1 A web-based social tagging system.

high participatory nature urged users in adapting them as a form of information organization and exchange of content and experience with other users.

Here, we focus on these systems that facilitate the storage and sharing of multimedia content. Currently, millions of users participate in multimedia social tagging systems, uploading content, adding tags or just browsing for tracking interesting content. The increased popularity of such sites can be traced by a rapidly increasing number of multimedia resources posted. Indicatively, we quote that YouTube reported in July 2006 100 million video viewings and 65 000 video uploads per day and Flickr is valued to have an upload rate of approximately 3000 images per minute, which yields 1.6 billion images per year. This realization is largely attributed to the widespread adoption of high quality but relatively low-cost digital media technology, which resulted in an enormous growth of readily available multimedia content.

Social tagging systems 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. As mentioned earlier, 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 [8].

This can be overcome by exploiting the annotations (tags) given in a multimedia social tagging system 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 many cases the concepts involved in the tags are ambiguous and there is subjectivity introduced by the users. Consequently the use of information extracted from visual features of the data can improve the accuracy of the method. Complementing the knowledge from tags with knowledge extracted from the content of the images is shown that can result in collecting valuable metadata that enhance the multimedia content exploitation.

3 State of the art

Currently, there is a growing number of research efforts that have focused on exploiting knowledge stored and often "hidden" in social tagging systems. However, in most of them, the resource management is a transparent process, which does not rely on the varying nature of digital resources (i.e. text or multimedia). Each resource is associated only with user-generated metadata (produced through the tagging activity), regardless of the specific nature of it. These involve: *i)* 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. In some approaches, though, analysis techniques for intrinsic feature extraction are employed, in order to achieve a better insight to the annotated content. Here, as outlined in the Introduction, we present an approach of web knowledge emergence, in which all tracks of knowledge (i.e. social, semantic and content-based) regarding the social content are taken into account. We give emphasis on multimedia content and especially on the knowledge that can be derived through low-level content-based multimedia analysis.

Towards that direction, the rest of this section is organized as follows. At first, a description of approaches that implement knowledge retrieval in social tagging systems, without employing content information is given. Then, a state-of-the-art on multimedia content-based related literature follows. The section ends with a presentation of approaches relevant to our technique in utilizing both tagging and content-based information for better retrieval.

3.1 Knowledge Retrieval in Social Tagging Systems

The dynamics of social tagging systems have turned a big part of scientific community into analyzing them and examining the emergent knowledge that derives from them. More specifically, in [3] and [2] the authors demonstrate that the structure and dynamics of social tagging systems are similar to those of a *complex system*, i.e. knowledge is built incrementally in an evolutionary and decentralized manner, yielding stable and knowledge-rich patterns, namely Emergent Semantics ([4]). Likewise in [9] the authors show that the tag proportions each resource receives crystallizes after about 100 annotations, attributing this behavior to users common background and their tendency for imitation of other users' tagging habits. They reach to this conclusion after examining and analyzing the tagging behavior in del.icio.us and identifying tagging patterns and kind of tags people tend to use.

Clustering is often introduced in the bibliography of social tagging systems as a way of overcoming the intrinsic limitations these systems have and, at the same time, generating knowledge from the mass activity. The authors in [10], [11] and [12] rely solely on tagging information and tag co-occurrence to derive semantically-related groups of tags and resources, out of social tagging systems. Each group of tags involves a certain topic and encompasses the users' understanding and vocab-

ulary describing this topic. Flickr photo-sharing system implements tag clusters, based on tag co-occurrence, as well, and handles quite well the tag ambiguity issue, managing to separate different senses of ambiguous tags in different cluster. For instance the ambiguous tag "*jaguar*" yields three clusters. The first cluster contains images and tags that describe the animal, the second one involves car-related material, while the last one includes tags and photos related to music. However, the described 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-occurring tags, which cannot be mapped to an actual topic and cannot be interpreted by a user. Additionally, they do not always tackle quite well the tag synonymy issue, since synonymous tags are commonly given by different users and they seldom co-occur.

To address the problem of lack of relations and semantics in the tag space, many researchers claim that the application of mature semantic web technologies (e.g. ontology usage, reasoning) on social data could add great value to the latter, as it may render a kind of structure to them. More specifically, in [13], the author proposes the building of an ontology that formalizes the activity of tagging, so to as enable the exchange, comparison and reasoning over the tag data acquired from varied tagging applications. Likewise, in [14] and in [15] the authors present their own OWL ontologies that aim at achieving a common formal conceptualization for the representation of tagging. Moreover, a step towards semantics' inclusion in a tagging system is the use of Simple Knowledge Organization System (SKOS) vocabulary [16], which allows to declare relationships between the terms used by users (e.g. broader term, narrower term, etc). Despite the fact that interoperability between tagging systems is a subject of research, these approaches have not found widespread application and, so far, there is no common agreement on a formal representation of tagging activity between social tagging systems.

Another trend for social data exploitation is the exploration of the tag space and the detection of emergent relations in social data that can be exploited for ontology building and/or evolution. It is expected that merging the Semantic Web with natural language and concepts used by ordinary people is a right step in the direction of making Semantic Web dynamic and bridging the gap between knowledge applications and common users. Towards that direction Schmitz, in [17], analyzes a model that employs natural language processing techniques to induce an ontology from Flickr tags. In [18], Mika proposes a model to extend the traditional bipartite model of ontologies with the social context in which each concept or instance is produced. He extracts community-based ontologies or evolves defined ones, based on emergent semantics from the underlying social tagging activities and claims that when social actions of a community are taken into consideration, the extracted ontology has greater potential to closely match the conceptualization of the corresponding community. Another approach of eliminating the lack of semantics in tagging systems can be found in [19], where the authors employ association rule mining, in order to analyze and structure the tag space. Likewise, they use the mining results for ontology learning. In [20], the authors try to tackle the shortcomings of a tagging system and extract semantics by clustering of tag data based on co-occurrence

and mapping of tags to ontology concepts with the use of semantic web engines. In the same way in [21], the authors use statistical analysis of co-occurrence of objects (in unsupervised learning, i.e. clustering) to infer a global semantic model. This semantic model can help in tag disambiguation and attempts to tackle the synonymy problem by grouping synonymous tags together. Finally, in [22] Zhou et al. present a clustering method for exploring hierarchical relations in social data.

The aforementioned overview of existing approaches indicates that clustering is, quite often, employed as a technique to overcome the limitations and improve the retrieval efficiency of social tagging systems.

3.2 Content-based Multimedia Retrieval

Multimedia information has replaced in recent years the traditional forms of storing knowledge, as printed text or still graphics. A wide variety of content forms is used nowadays: text, audio, still images, animation, video, interactivity. Consequently methods for multimedia retrieval and mining are necessary for the effective use of multimedia information. Content-based multimedia analysis is necessary, because even though text in many cases is present, it is ambiguous. In addition, there is subjectivity introduced by the human annotator. While both visual, audio and other content-related features can be used in content-based methods to improve retrieval accuracy, in this work we focus on the use of visual information. Lew et al have made an excellent work in gathering all research trends in their survey paper [23] where they also pinpoint what is to be expected from new research efforts in the field.

New features and similarity measures are proposed and used in order to efficiently describe multimedia information and consequently help to fulfill the goals of multimedia information retrieval. MPEG-7 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) that standardizes object-based audiovisual description tools, including the metadata elements and their structure and relationships that create descriptions enabling effective and efficient access to multimedia content. MPEG-7 allows fast and efficient content searching, filtering and identification, and addresses a large range of applications [24].

Lew [25] and Gevers [26] propose new color features that are applied in fields such as lighting invariance, intuitiveness and perceptual uniformity. Research on texture understanding has been done by Ojala et al. in [27] that outline the effectiveness of using simple texture histograms. Additionally a new texture feature based on the Radon transform orientation is introduced in [28]. Novel approaches on learning shape have been proposed in [29, 30, 31]. In [32] Vretos et al. propose several classes to extend the MPEG-7 standard and describe the digital video content in a more homogeneous and anthropocentric way.

In content-based multimedia retrieval, similarity measures have an equally important role as the visual features. In [33] Sebe et al. provide a method for selecting the appropriate metric given a training dataset and propose the Cauchy metric as

an alternative to the commonly used distance measures. Jacobs et al. [34] evaluate the performance of nonmetric distances in classification. New methods of measuring image similarity based on graph matching and time and pictorial content are suggested in [35] and [36] respectively.

Lindeberg [37] presents a scale selection methodology, using the Laplacian-of-Gaussian function. The computation of the size of image structures can be done from the scales at which normalized differential geometric descriptors assume maxima over scales. Scale-invariant feature transform (SIFT) [38] is an algorithm in computer vision to detect and describe local features in images. This algorithm in the first step constructs a scale space pyramid using difference-of-Gaussian filters. The Laplacian-of-Gaussian can be approximated using the difference-of-Gaussian. From the local 3D maxima, a robust descriptor is built for matching purposes. The localization of the features that are detected using difference-of-Gaussian and Laplacian-of-Gaussian may not be very accurate. This disadvantage is due to the fact that they respond to high gradients and consequently the repeatability is not the best possible.

In the field of evaluation, TRECVID [39] has been the most complete evaluation initiative during the last decade and has benchmarked detection of a variety of semantic and low-level video features. Additionally, in recent years, there has been an extended utilization of explicit knowledge with formal semantics which within the SW initiative translates to the use of ontologies [40, 41, 42].

3.3 Integrating Social with Content-based Knowledge

Recently, there has been an increasing interest by the research community towards approaches that utilize both tagging information and content-based features.

In [43], the authors claim that the intrinsic shortcomings of collaborative tagging are tackled by employing content-based image retrieval technique. The user is facilitated in image database browsing and retrieval by exploiting both the two aforementioned technologies in a supplementary way. Indeed, it is shown that the visual features can support the suggestion of new tags and contribute to the emergence of interesting (semantic) relationships between data sources. Through the use of a navigation map, these emergent relationships between users, tags and data may be explored. The visual features employed for the content-based image retrieval are *Color* and *Texture*. For the extraction of texture features they use *Oriented Gaussian Derivatives*[44].

Our original approach for coupling tagging information with content-based features was introduced in [45]. A number of varied clustering techniques were employed and applied to a dataset from Flickr. The clustering was tag-oriented and occurred in two steps. In the first step the resources were assigned to clusters, depending on the similarity of their accompanying tags. The similarity between tags yields based on their co-occurrence in tagging activities of users and their semantic vicinity. For every cluster an emergent topic was extracted based on the most frequent tags used to describe the resources assigned to this cluster. In the second step,

visual features were employed, in an effort to increase the purity of already created clusters. For instance, if an image assigned to the cluster “sea” was found quite dissimilar to the rest images, it was removed from the specified cluster as an outlier. The second step of the process could be regarded as a “*misleading tags tracking phase*”. The evaluation showed that the resulted clusters were very good, each one containing images and tags about the topic it has been extracted from the specified cluster. This approach was extended and presented in the next section of the chapter. Another work that combines user data with feature-based approaches, in order to rank the results of a video retrieval system is presented in [46]. The authors use this knowledge, along with a multimedia ontology to build a learning personalized environment.

A number of works have addressed the problem of identifying photos from social tagging systems that depict a certain object, location or event [47, 48, 49]. In [47] they analyze location and time information from geotagged photos from Flickr, in order to track tags that have place semantics (i.e. they refer to an object in a restricted location) or event semantics (i.e. they are met in specified time periods). Then, they employ tag-based clustering on these specific tags, followed by visual clustering, in order to capture distinct viewpoints of the object of interest. The same authors in [50] combine tags with content analysis techniques, in order to get groups of music events photos. Likewise, in [48, 49] the authors use various modalities of photos (i.e. visual, textual, spatial, temporal proximity), in order to get photo collections in an unsupervised fashion. Apart from the obvious retrieval application, the outcome of these methods can be used for training of multimedia algorithms and for tag recommendations. Another approach towards this direction, that deploys the visual annotations, also known as “notes” in Flickr is described in [51], where it is shown that the retrieval of content in a social tagging system improves significantly by combining tags and visual analysis techniques.

The problem of tag recommendation has been studied in [52], where the authors suggest an approach for recommending tags by analyzing existent tags, visual context and user context in a multimedia social tagging system. Tag recommendation techniques were, also, proposed in [53], where the authors suggest four methods for ranking candidate tags and in addition, they present the semantics of tags in flickr.

Other efforts to design tools that employ simple image analysis algorithms and apply them on Flickr images have appeared in [54], [55], yet they are not intended for semantic similarity extraction or integrated navigation in the social tagging system.

4 Content and Tag-based Clustering Approach

In this section we present a two-step method for clustering on multimedia social sources. As highlighted in section 3.1, clustering is often introduced in the bibliography of social tagging systems as an approach to overcome their intrinsic limitations and derive knowledge regarding their content or their users. The main approach is:

divide the resources into semantically related clusters (i.e. meaningful groups of resources) and exploit the shared understanding about tags and resources fostered in each cluster. The division is performed according to some *metric of similarity* and each extracted cluster would ideally correspond to a specific topic. The expected benefit of the whole process is that the collective activity of tagging will isolate erroneous tags and illustrate the dominant tags in each cluster, expressing, thus, the community's point of view around the corresponding topic.

In order for the clustering to be effective and yield pure clusters, an appropriate metric of similarity between the resources needs to be employed. In an effort to capture knowledge in all its forms, a *two-step process* is adopted. In the *first step* the textual knowledge about the resources is considered. This involves capturing social and semantic similarity of the resources' accompanying tags. The intuition here is that if the similarity among the tags of two resources is high, then the resources are possible related to one another. In the *second step* of the process, content-based methods are employed, so as to get additional insight into the multimedia content. While both visual, audio and other content-related features can be used in content-based methods to improve retrieval accuracy, in this work we focus on the use of visual information.

Based on these, the rest of this section is organized as follows. At first, a problem formulation is quoted, to emphasize the required concept definitions and the mathematical notations used throughout the rest of the chapter. Then, an analytical description of each step of the process follows.

4.1 Problem Formulation

We define a Social Tagging System as the finite sets U, R, T, A which describe the set of users, resources, tags and user annotations (i.e. tag assignments), respectively. Table 1 summarizes the basic symbols' notation used in this paper.

Table 1 Main Symbols' Notation

Symbol	Definition
m, n, l	Number of users, resources, tags (respectively)
d, p	Number of attributes and user annotations (respectively)
K	Number of clusters
U	Users' Set $\{u_1, ..., u_m\}$
R	Resources' Set $\{r_1, ..., r_n\}$
U	Tags' Set $\{t_1, ..., t_l\}$
A	User Annotations' Set $\{a_1, ..., a_p\}$
AS	Attributes' Set $\{at_1, ..., at_d\}$
MA	Manual Annotations' Set $\{ma_{r_1}, ..., ma_{r_n}\}$

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, as follows:

Definition 1 (RESOURCE'S REPRESENTATION**).** Each resource $r_j \in R$, where $j = 1 \ldots n$, is represented by aggregating the tags assigned to it by all users. Thus:

$$r_j = (h_1 \times tag_{j1}, h_2 \times tag_{j2}, \ldots, h_z \times tag_{jz}) \quad (1)$$

where z is the number of tags assigned to r_j by all users and the coefficients $h_i, i = 1, \ldots, z$ denote the number of times the tag_{ji} has been used in r_j's annotation.

Our purpose is to create groups of related resources by taking into consideration textual annotations and content-based information and, thus, we need to provide solution to the following JOINT SOCIAL, SEMANTIC & CONTENT-BASED DATA CLUSTERING problem.

Problem 1 (JOINT SOCIAL, SEMANTIC & CONTENT-BASED DATA CLUSTERING**).** Given a set R of n resources, an integer k and a *Similarity* function, find a set C of k subsets of resources, $C = \{C_1, \ldots, C_k\}$, such that $\sum_{x=1}^{k} \sum_{r_i, r_j \in C_x} Similarity(r_i, r_j)$, $i, j = 1, \ldots, n$ and $i \neq j$, is maximized.

The *Similarity* function must be defined in a way to sufficiently capture the association between two resources by jointly considering the social and semantic aspects of their accompanying tags, together with the low-level visual features of the involved resources. These two types of data have very different characteristics: while textual data (i.e. tags) is typically sparse and high-dimensional, visual data is usually dense and low-dimensional. Due to this heterogenous representation of the two modalities involved in the feature space, a *two-step* process is followed, in each step of which, each modality is treated separately.

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

Definition 2 (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 3 (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.

The two steps of the proposed framework are shown in Figure 2.

4.2 Tag-based Resources Clustering

This section describes the first step of the proposed method, which aims at a tag-guided resources' clustering. As already discussed in section 4.1, in our approach,

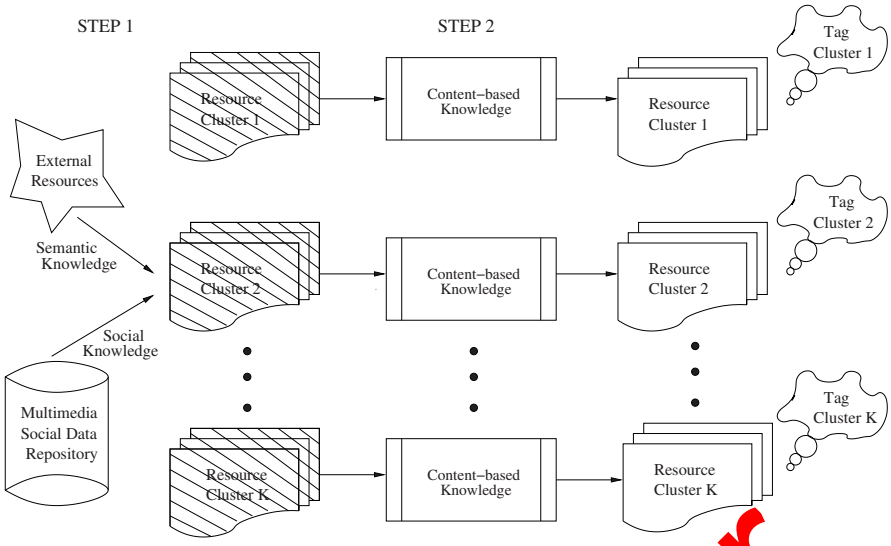

Fig. 2 Two-step content and tag-based clustering approach.

each resource is expressed via the tags assigned to it (see equation 1). In practice, the number of tags used to represent all the resources in a social tagging system may grow in large scale and thus we need to employ a selection process of the most distinguishing tags which will form the resources' attribute set *AS*. In our approach we use the d most frequent tags to form the *AS* set which will guide our clustering process.

Definition 4 (THE ATTRIBUTE SET**).** Given the $T = \{t_1, \ldots, t_l\}$ set of tags, we define the attribute set $AS = \{at_1, \ldots at_d\}$: $AS \subseteq T$ and *AS* contains the d most frequent tags $t_x \in T$.

Each attribute $at_y \in AS$ is related with a different degree to the various r_i, $1 \leq i \leq n$, resources, while two different resources may be indirectly related, if they present strong relation with the same set of attributes. To represent the relation of each resource to each attribute, we define a function, namely *SimilarityFactor* sf_{ij} between a resource r_i and an attribute $attr_j$ that is evaluated by encompassing both social and semantic similarity between the resource's tags and the tag that corresponds to the specified attribute. We describe in the sequel how this similarity yields.

As introduced in section 3.1, current approaches which employ clustering in social tagging systems, rely solely on tag co-occurrence to estimate tag closeness, and hence, resource closeness. We refer to such similarity between two tags as *social similarity*, *SoS*, and we define it as follows:

$$SoS(t_x, t_y) = \frac{\sum_{i=1}^{n} r_i : (u_w, r_i, t_x) \in A \; and \; (u_z, r_i, t_y) \in A}{\max(\sum_{i=1}^{n} r_i : (u_w, r_i, t_x) \in A, \sum_{i=1}^{n} r_i : (u_z, r_i, t_y) \in A)} \quad (2)$$

where $u_w, u_j \in U$, $r_i \in R$.

However, considering the semantic aspect of tags, as well, is expected to be beneficial for the clustering process in a social tagging system, since it can contribute to eliminating the tag synonymy issue and avoiding separation of semantically related tags into different clusters. For the estimation of the *Semantic Similarity* between two tags, we need to use external resources (i.e. web ontologies, thesauri, etc), available in the web. 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 thorough presentation of the most popular ones is given in [56]. In our work we adopted the Wu & Palmer measure, described in [57], due to its straightforward application to our data. According to this measure, the semantic distance between two concepts is proportional to the path distance between them. For example, let t_x and t_y be two tags for which we want to find the semantic similarity and $\overrightarrow{t_x}$, $\overrightarrow{t_y}$ be their corresponding mapping concepts via an ontology. Then, their *Semantic Similarity SeS* is calculated as:

$$SeS(t_x, t_y) = \frac{2 \times depth(LCS)}{[depth(\overrightarrow{t_x}) + depth(\overrightarrow{t_y})]} \tag{3}$$

where $depth(\overrightarrow{t_x})$ is the maximum path length from the root to $\overrightarrow{t_x}$ and LCS is the least common subsumer of $\overrightarrow{t_x}$ and $\overrightarrow{t_y}$.

The total similarity between two tags will be estimated by considering both their social and semantic similarity, which are normalized in the interval $[0..1]$ (Equations 2, 3). In order to examine the impact that each kind of information has on the clustering process, we combine 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 two tags t_x and t_y in terms of both their social (Equation 2) and semantic (Equation 3) similarity as:

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

where $w \in [0, 1]$ is a normalization parameter which adjusts the magnitude of the semantic similarity against the social one upon the final outcome. More specifically, at the one end when $w = 1$ we consider solely the *Social Similaty SoS*, while at the other end, when $w = 0$, only the *Semantic Similarity SeS* is considered. For any other value of w both similarities contribute to the *Similarity Score SS* of two tags.

Having specified the similarity metric between tags, we can proceed to the estimation of $similarity factors, sf_ij$, discussed in the beginning of the section.

Definition 5 (SIMILARITY FACTOR**).** Given a resource r_i, in which the users have assigned $|r_i|$ tags, and an attribute $attr_j$, we define as Similarity Factor, sf_{ij}, between the specified resource and the specified attribute, the maximum Similarity Score, SS between every tag assigned to resource r_i and the attribute $attr_j$. Thus:

$$sf(r_i, attr_j) = max_{x=1...|r_i|}\{SS(t_x, at_j)\} \tag{5}$$

where $r_i \in R$, $t_x \in r_i$, $at_j \in AS$.

In the above definition, we assume that all the tags assigned to each resource are relevant to the content. Alternatively, taking the average *Similarity Score* 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 (where 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:

$$SimMatrix(i, j) = sf(r_i, at_j) \tag{6}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, d$.

The above resources *similarity matrix* is the input to the clustering procedure, out of which k resources clusters shall arise. As described above, the *similarity function* that is used to estimate the relation between two resources (in this phase) is based on both social and semantic aspects of their involving tags.

4.3 Cluster Refinement with MPEG-7 Visual Features

This section describes the second step of the proposed approach, which involves clustering of the resources, based on their visual features. Content-based approaches are often employed in the multimedia content retrieval, as can be seen in the bibliography presented in section 3.2. Here, we exploit multimedia analysis as a means that gives additional insight into the content (apart from the present textual annotations) and is expected to minimize the intrinsic limitations of social tagging systems and potentially improve the retrieval accuracy.

In order to estimate the visual 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 [58]. The MPEG-7 standard constitutes the greatest effort towards a common framework to multimedia description. It aims to provide a rich set of standardized tools for the description of multimedia content and additionally support some degree of interpretation of the information's meaning enabling thus smooth sharing and communication of multimedia metadata across applications and their efficient management, e.g. in terms of search and retrieval. MPEG-7 is implemented in the form of XML Schemas.

The MPEG-7 Standard consists of five main parts: the *Description Definition Language* that defines the syntax of the MPEG-7 Description Tools and new Description Schemes, the *Visual* and *Audio* parts that include the description tools for visual and audio content respectively, the *Multimedia Description Schemes* that

comprise the set of Description Tools dealing with generic features and multimedia descriptions and the MPEG-7 Systems, the tools needed to prepare MPEG-7 descriptions for efficient transport and storage and the terminal architecture.

The MPEG-7 Visual Description Tools, that are included in the standard and are related to our approach, consist of basic structures and descriptors that cover the following basic visual features: color, texture, shape, motion, localization, and face recognition. In Table 2 there are the visual features and their corresponding MPEG-7 visual descriptors.

Table 2 MPEG-7 descriptors of visual features

Color	Texture	Shape	Motion
Color Quantization	Texture browsing	Region Shape	Motion Activity
Dominant Color	Edge histogram	Contour Shape	GoF/GoP
Scalable Color	Homogeneous Texture		
Color Layout			
Color Structure			

Color and texture descriptors are among the most expressive visual features. This is the reason why they are widely used and they were chosen in our own case in order to extract visual information from the images. In particular we used three Color Descriptors of MPEG-7: Scalable Color, Color Structure, Color Layout and two Texture Descriptors of MPEG-7: Homogeneous Texture and Edge Histogram [59]. An extended description of the five MPEG-7 descriptors that we used can be found in the Appendix 8.

MPEG-7 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 [60].

Therefore the second step of our approach is based on identifying low-level visual 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 the descriptors of the MPEG-7 standard, mentioned above 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.

5 Experimental results

In this section, experimental results of the application of the proposed approach to a corpus of multimedia resources obtained from a social tagging system are presented.

To carry out the experimentation phase and the evaluation of the proposed clustering approach, two different datasets from Flickr were crawled using the wget[5] utility and Flickr API facilities. The first one consists of 3000 images depicting cityscape, seaside, mountain, roadside, landscape, sport-side and locations (about 500 images from each domain). The second dataset comprises 10000 images related to concepts: jaguar, turkey, apple, bush, sea, city, vegetation, roadside, rock, tennis. The particular selection was based on the fact that the above concepts are very commonly used by Flickr users and embed ambiguity that restricts their efficient retrieval. As a source of semantic information for tag concepts, we employ the lexicon WordNet [61], which stores English words organized in hierarchies, depending on their cognitive meaning.

Both image datasets were manually annotated in order to get the ground truth for the evaluation procedure. Even though manual annotation of 13000 images is a big task both time consuming and tedious, it enables the testing of our method, using quantitative measures (like precision, recall and f-measure), rather than relying solely on qualitative observation of the data or on (often misleading) user tags. In addition, the gathered dataset together with the manual annotations is a valuable source for the training of multimedia analysis algorithms. Next, we describe the metrics that we used to evaluate our proposed approach.

5.1 Evaluation Metrics

To evaluate the quality of the extracted clusters of resources, for each technique described in the chapter, each image resource was manually annotated with respect to a predefined vocabulary V related to the visual and thematic content of the images. Thus, the *Manual Annotations Set* was created, which contains the manual annotations, each resource has received, i.e.:

$$MA = \{\cup ma_{r_x}\}, \forall r_x \in R \tag{7}$$

Then, we use precision Pr and recall R as follows. Let, C_j be an extracted cluster and CT_j be the dominant tags assigned to resources of the specified cluster, above a user-defined threshold τ (see definition 3 - CLUSTER TOPIC). We call *Relevant Resources RR* of the cluster C_j the set of resources in the corpus that at least one of the manual annotations they have received matches a tag in CT_j, i.e.:

$$RR(C_j) = \{\cup r_x\}, \forall r_x \in R : ma_{r_x} \cap CT_j \neq \emptyset \tag{8}$$

[5] wget: *http://www.gnu.org/software/wget*

where $ma_{r_x} \in MA$.

It should be noted that in case we perform visual clustering in a tag-based cluster (this happens during the 2nd step of our proposed method), the *RR* are computed on the resources in the tag-based cluster, and not in the entire dataset. Thus, if C_j is a tag-based cluster and VC_i is a visual cluster extracted from C_j, then the *RR* of the VC_i are the set of resources in the C_j that at least one of the manual annotations they have received matches a tag in the cluster topic of VC_i, i.e.:

$$RR(VC_i) = \{\cup r_x\}, \forall r_x \in C_j : ma_{r_x} \cap CT_i \neq \emptyset \tag{9}$$

where $ma_{r_x} \in MA$.

We define *precision* as the fraction of resources that belong to C_j and are also relevant resources:

$$Pr(C_j) = \frac{|C_j \cap RR(C_j)|}{|C_j|} \tag{10}$$

We define *recall* as the fraction of relevant resources which belong to C_j:

$$R(C_j) = \frac{|RR(C_j) \cap C_j|}{|RR(C_j)|} \tag{11}$$

The ideas of *precision* and *recall* are combined in *F-Measure* which is a broadly accepted and reliable index used in various clustering evaluation approaches [63]. Given the precision and recall definitions described in this section, the value of F-measure for a cluster C_j is defined as:

$$F(C_j) = \frac{2 * Pr(C_j) * R(C_j)}{Pr(C_j) + R(C_j)} \tag{12}$$

The values of F-measure fluctuate in the interval $[0..1]$ with higher values indicating a better clustering.

The user-defined threshold τ sets the frequency limit a tag should reach, in order to be member of the *Cluster Topic* of the specified cluster. It takes values in $[0..1]$, where $\tau = 1$ denotes that a tag should have been assigned in every resource in the cluster, so as to be part of the *Cluster Topic*, while $\tau = 0$ denotes that all the tags assigned to cluster resources are also members of the *Cluster Topic*. After testing varying values for τ, we concluded that the best value for the specified dataset was 0.6 (i.e. the *Cluster Topic* of every extracted cluster comprises tags that have been assigned to at least 60% resources of the specified cluster, as shown in Figure 3).

5.2 Clustering Evaluation

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

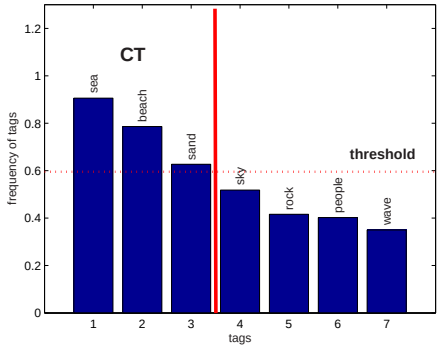

Fig. 3 Selection of τ value

hierarchical (agglomerative) and a conceptual clustering process (Cobweb) [64]. In the second step of the process we conducted experiments using content-based information of the images. For all the images in both datasets (13000 images) the low-level visual features were extracted. In order to remove the irrelevant images from each cluster we conducted experiments using different number and types of visual features. In particular, we evaluated the performance of each one of the 5 MPEG-7 Visual Descriptors mentioned in section 4.3 separately and the performance of every possible combination of groups of 2, 3, 4 and 5 Descriptors.

In Tables 3 and 4 the precision (Pr) and recall (R) of the clustering algorithms, as defined in the previous section (5.1), are quoted for different values of number of clusters that were extracted from the first dataset (3000 images), respectively. In each table, the measure is calculated at each step of the procedure 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 (average of 15% improvement).

Table 3 Precision in each step for varying algorithms and varying number of clusters (K) (1st dataset 3000 resources)

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*

Likewise, in Tables 5, 6, the precision and recall of each clustering algorithm on the second dataset (10000) are shown.

As it can be seen for K=20 the best clustering yields, for this specific dataset. For all algorithms the precision was satisfying, meaning that the extracted clusters were

Table 4 Recall in each step for varying algorithms and varying number of clusters (K) (1st dataset 3000 resources)

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	0.6
Cobweb	0.749	0.539	0.805	0.78	0.78	0.732
	Step 1	*Step 2*	*Step 1*	*Step 2*	*Step 1*	*Step 2*

Table 5 Precision in each step for varying clustering algorithm and varying number of clusters (K) (2nd dataset 10000 resources)

	K=10		K=20		K=30	
K-means	0.57	0.604	0.644	0.738	0.655	0.685
Hierarchical	0.801	0.497	0.542	0.78	0.693	0.764
Cobweb	0.71	0.712	0.7	0.7	0.696	0.7
	Step 1	*Step 2*	*Step 1*	*Step 2*	*Step 1*	*Step 2*

Table 6 Recall in each step for varying clustering algorithm and varying number of clusters (K) (2nd dataset 10000 resources)

	K=10		K=20		K=30	
K-means	0.4	0.489	0.456	0.531	0.245	0.307
Hierarchical	0.2	0.223	0.121	0.354	0.025	0.482
Cobweb	0.04	0.299	0.388	0.359	0.383	0.41
	Step 1	*Step 2*	*Step 1*	*Step 2*	*Step 1*	*Step 2*

of good quality. It is amazing that the precision of the clusters on 10000 images is on the same levels with the precision of the clustering on the smaller dataset. This holds for all the three clustering algorithms we tested and proves the scalability of our approach in extracting clean clusters. The low values of the recall are attributed to the big size of the dataset and they show that the proposed approach did not manage to capture all the relevant resources together. Finally, and in this dataset, in most cases there was an improvement from combining tag analysis with visual knowledge from the content.

It should be noted that all algorithms were applied for a certain number of times on our data (in order to avoid random assignments of data) and here we report the average performance.

5.3 Emergent Tag Clusters and Cluster Topics

Generally, most of the clusters the system generated were homogeneous and meaningful. The corresponding tag clusters were also very representative and highly informative. Indicatively F-Measure metric is presented for ten extracted clusters of

the dataset of 10000 images, along with the dominant tags for each cluster (i.e. its cluster topic). The values of F-measure fluctuate in the interval [0..1] with higher values indicating a better clustering.

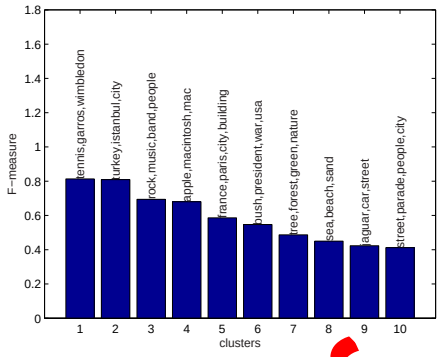

Fig. 4 Clusters' F-measure - 10000 dataset, 10 clusters.

5.4 Influence of w in the extracted clusters

In Section 4.2, the similarity between two tags was defined as a weighted sum of their social and semantic similarity (Equation 4). The parameter w takes values in $[0,1]$ and is used to adjust the impact each kind of similarity measure (i.e. social or semantic) has on the overall outcome. More specifically, when w is close to 1, the social similarity is favored, while when w approximates 0, the semantic similarity is mostly considered in the total tag similarity calculation.

Here, we will examine how the values w takes, affects the quality of the extracted clusters. We will experiment with the following 3 indicative cases:

- $w = 0.2$: The similarity between 2 tags is mostly based on their co-occurence.
- $w = 0.5$: Both co-occurence and semantic affinity between 2 tags are counted equally in the estimation of their similarity.
- $w = 0.8$: The similarity between 2 tags is mostly based on their semantic affinity.

The F-measure of four indicative clusters for each value of w is shown in Figures 5, 6 and 7, respectively. The specified clusters were obtained with the hierarchical algorithm. It should be noted that these clusters are tag-based clusters (obtained during the first step of our proposed approach), since the value w affects the way we calculate tag affinity. The effect on the second step is indirect: that is, the better clusters yield during the first step, the higher the improvement in the overall procedure.

As can been seen in all figures, the value of w affects the results. More specifically, we observe that, in most cases, for $w = 0.5$ both precision and recall have their

highest values, meaning that the incorporation of both kinds of knowledge (social and semantic) is more advantageous towards relying solely on one of them.

In case of $w = 0.2$, where more weight is given to the *Social Similarity*, we can derive that the objects assigned by the algorithm in the same cluster have tags that co-occur in the users' annotations. For example the tags *forest*, *nature*, *green*, *tree* belong to the same cluster, because these tags are often used together for describing images related to sceneries of nature. The same holds for the cluster where *street*, *building*, *church*, *architecture* are assigned, since they constitute tags that occur frequently in the description of images referring to city places. In general, tag co-occurence has proven to be more advantageous in the case of ambiguous tags (homonyms), since it is the context of such a tag (i.e. its co-occuring tags) that will help to disambiguate its meaning. However, lacking semantic information, the algorithm splits meaningful clusters into subclusters. This explains the low recall in the *sea* and *vegetation* clusters, in Figure 5.

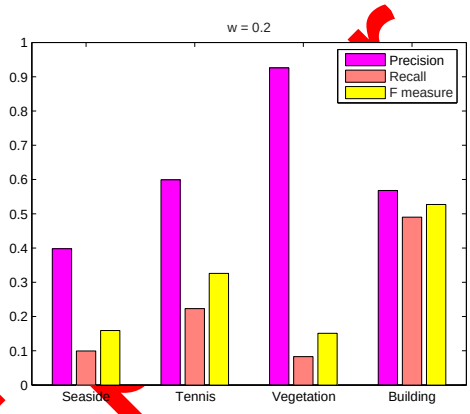

Fig. 5 F-measure of 4 clusters, taken from hierarchical algorithm with w=0.2.

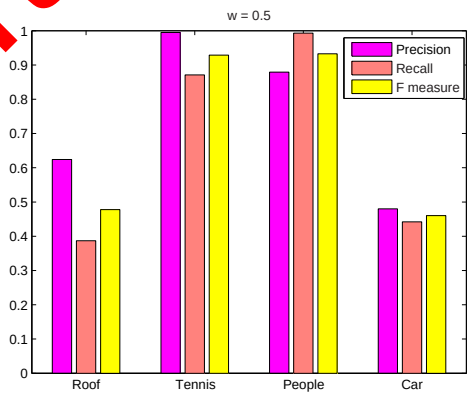

Fig. 6 F-measure of 4 clusters, taken from hierarchical algorithm with w=0.5.

For $w = 0.8$, (Figure 7), where the *Semantic Similarity* is favored, the algorithm assigns all semantically close tags in one cluster i.e. *sea*, *seaside*, *beach*, *sand* (*beach* and *sand* are grouped together). Despite the fact that all aforementioned tags are closely akin, in the previous described cases, they are split into different clusters, due to the fact that the users have not used all of them together in their annotations. However this method fails in disambiguating correctly tags like *rock* and *rocks*, to the same cluster even though in most cases they are not used in the same sense and they do not describe the same set of images. This results in clustering images having the tag *rock* but involving music themes together with images depicting stones. Thus, we can conclude that while this approach yields semantically meaningful clusters around a specific topic and it tackles well in case of synonyms (or tags with alike meaning), it fails to handle the tag ambiguity issue.

6 Use Cases - Scenarios

In this section we will show some indicative use cases and scenarios of our proposed approach. First of all, the proposed method tackles quite well the shortcomings of a social tagging system, described in the Introduction, resulting, thus, in better retrieval of multimedia content. Furthermore, the extracted clusters together with the cluster topics can be used as training sets for multimedia analysis learning algorithms [62]. Apart from the multimedia related application, our method has an ability in subdomain identification within a domain, which can be utilized in semantics extraction out of the raw tag data. Another potential use case of our method would be its integration by a recommender system, in order support users of tagging systems by suggesting tags that have been already assigned to related content. In the following, some scenarios are demonstrated that justify our arguments. Due to space restriction only some snapshots are shown indicatively.

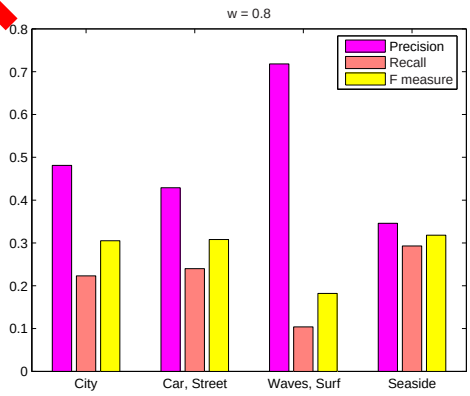

Fig. 7 F-measure of 4 clusters, taken from hierarchical algorithm with w=0.8.

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, by adjusting accordingly the value of *w*, as explained in the previous section (see Figure 8).

Tag 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, in cases where a misleading tag may lead to the retrieval of irrelevant content, then the content similarity factor, employed in step 2 of the process, enhances the possibility that irrelevant content will be tracked and removed from a cluster, if the referred object has a visual appearance very different from the rest ones (e.g. Figure 9). The cluster shown is a snapshot of the outcome of step 1 and the resource surrounded by a red box is removed during the step 2 of the process. The removed photo has been assigned with the tag ”sea” which was a misleading tag and it was tracked.

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 (Semantics extraction): The proposed approach 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 Figure 10, with (a) CT = building, roof, street, (b) CT = car, race, Porsche, street (c)CT = caribbean, carnival, festival, people, street.

Tag recommendation: The emergent cluster topic of each cluster can be suggested as candidate tags for the objects assigned inside the cluster. Furthermore,

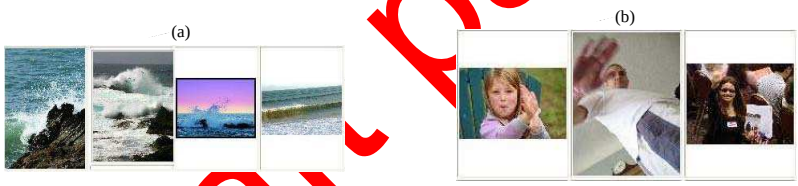

Fig. 8 Different clusters for the ambiguous tag: wave (a) members of cluster with CT = wave, sea, water (b) members of cluster with CT = wave, signal, hand, person(The photos are downloaded from Flickr photo sharing system)

Fig. 9 Snapshot of a sea cluster with its emergent CT - Identification of a misleading tag in the sea cluster and rejection of the resource (surrounded by a red box) (The photos are downloaded from Flickr photo sharing system)

ranking mechanisms for candidate tags can be developed, based on the visual similarity of the content.

7 Conclusions

This paper introduces a joint approach for social data grouping that aims to enhance the multimedia social content exploitation. The proposed method considers the semantic in addition to the social aspect of resources accompanying tags in a balanced way, as well as the content-based information. It yields clusters consisting of both resources and user annotation tags. The proposed approach has been evaluated under two real datasets and the results proved its efficiency in extracting relevant tags and resources, illustrating the dominant tags in each cluster and expressing users' point of view around the corresponding topic. Moreover, the consideration of the visual aspect of the social resources enables the satisfying handling of common social tagging limitations, such as the tag ambiguity issue. The proposed approach 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. In addition, the calculation of the similarities were relatively time consuming, so we plan to study ways to decrease the time spent and experiment with different metrics.

Fig. 10 Members of different clusters of roadside images (The photos are downloaded from Flickr photo sharing system)

Acknowledgements The work presented in this paper was partially supported by the European Commission under contracts FP7-215453 WeKnowIt and FP6-26978 X-media.

8 Appendix

8.1 Color Descriptors

(1) Scalable Color

The Scalable Color Descriptor (SCD) is defined in the HSV color space (see HSV Color Space description below). It uses an encoding method, based on Haar transform on the color HSV histogram. The HSV space is uniformly quantized into histogram bins. The number of the bins can vary. The number of the bins depends on the required compactness- a low number of bins give a fast descriptor suitable for indexing and quick queries.

After the histogram values are extracted, there is a normalization and a nonlinear quantization into a four-bin integer representation. The Haar transform is applied to the four-bit integer values across the histogram bins. The output is the high and low-pass coefficients from the transform. In Figure 11 it is described the process of SCD extraction process.

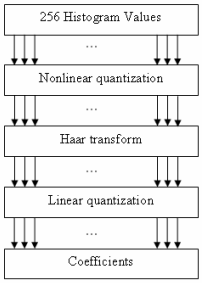

Fig. 11 Diagram of SCD generation

Haar transformation creates a scalable description, which is useful for image-to-image matching and retrieval based on color feature. In addition, SCD can be further used for group of frames or groups of pictures in video data.

(2) Color Structure

The Color Structure Descriptor (CSD) represents the local color structure in an image. This descriptor enables it to distinguish both in which proportion each color

exists and how uniformly color is distributed in the image. The CSD is a histogram and is computed by the use of an 8x8 structuring element, which visits all location in the image as shown in Figure 12. In particular, CSD counts how many times a particular color is contained in all the pixels in the 8x8 window, as this window scans the image. Suppose $C_0, C_1, C_2, \ldots, C_{M-1}$ denote the M quantized colors. The value of each bin of the histogram represents the number of the structuring elements in the image that contain one or more pixels with the corresponding color Cm. In this way, unlike the color histogram, the CSD can help us to distinguish two pictures with the same amounts of color but with different color distribution. The CSD uses the HMMD color space, which is quantized non-uniformly into 32, 64, 128 or 256 bins. In our case the number of bins is 32. An 8-bit code represents each bin amplitude value.

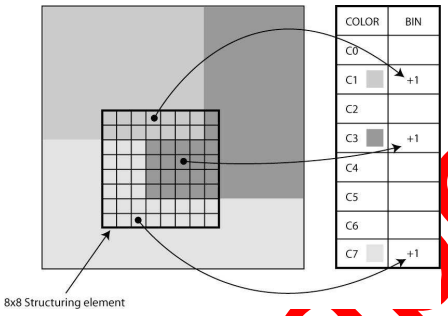

Fig. 12 CSD structuring elememt [65]

As it is mentioned above, the size of the structuring element is 8x8. Their number is always 64, consequently the distance between the structuring points increases with the image size, as shown on Figure 12. The spatial extent of the structuring element is computed by the following rule:

$$p = \max\{0, round(0.5 \log WH - 8)\} \tag{13}$$

$$K = 2^p, E = 8K \tag{14}$$

where

W, H image width and height;
$E \times E$ spatial extent of the structuring elements;
K sub-sampling factor.

For images smaller than 320 x 240pixels, an 8x8 structure element with no subsampling is used, and for image size 640 x 480 (p = 1, K = 2, and E = 16) structuring element is 16x16 and subsampling is 2x2. The structuring element of size 8x8 is applied to a subsampled image.

(3) Color Layout

The Color Layout Descriptor represents the spatial distribution of color of images in a very compact form. Its computation is based on the generation of an 8x8 thumbnail of the image. This 8x8 image is a result of DCT of the initial image and quantization. In particular the CLD extraction process consists of two parts. Firstly the input image is divided into 64 blocks (8×8). For each block of the grid, its average color is used as the representative color of the block. The average color is expressed in the YCbCr color space. An 8×8 DCT is performed in order to transform the derived average colors into a series of coefficients. After the transformation, the coefficients are zigzag scanned and few low-frequency coefficients are selected and quantized.

For matching two CLDs, $\{DY, DCr, DCb\}$ and $\{DY', DCr', DCb'\}$, the following distance measure is used:

$$D = \sqrt{\sum_i w_{yi}\left(DY_i - DY_i'\right)^2} + \sqrt{\sum_i w_{bi}\left(DCb_i - DCb_i'\right)^2} + \sqrt{\sum_i w_{ri}\left(DCr_i - DCr_i'\right)^2} \tag{15}$$

This descriptor provides an image-to-image or sketch-to-image search, which is high speed, accurate and requires minimum storage and transmission cost.

8.2 Texture Descriptors

(1) Homogeneous texture

The Homogeneous Texture Descriptor provides an accurate quantitative description of texture. The extraction method of the HTD is as follows: the frequency domain is partitioned into 30 channels as it is shown in Figure 13. The partitioning of the frequency space is uniform in the angular direction (step size of 30^o), but in the angular direction there is an unequal octave division. The individual feature channels are modeled by 2D-Gabor functions. The energy and the energy deviation of each channel is computed. Finally, mean and standard deviation of frequency coefficients are computed, resulting in a feature vector of 62 values as it is shown in equation 16. The HTD can be used for accurate search and retrieval.

$$TD = [f_{DC}, f_{SD}, e_1, e_2, \ldots, e_{30}, d_1, d_2, \ldots, d_{30}] \tag{16}$$

(2) Edge histogram

The Edge Histogram Descriptor represents the spatial distribution of five types of edges. In particular, the computation of the descriptor consists of four steps:

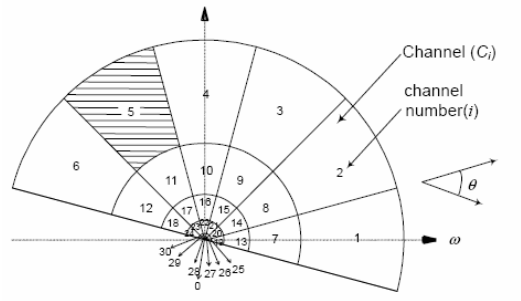

Fig. 13 Channels used in computing the HTD [66]

1. The image is divided in 4x4 sub-images.
2. Each of the sub-images that occur is divided in square blocks
3. Each block is described by one edge type. There are four directional edges (horizontal, vertical, diagonal 45^o, diagonal 135^o) and one non-directional.
4. The edge histogram is extracted

EHD represents local edge distribution in the image. However, the local histogram bins can be used in order to generate global and semi-local edge histograms, which increase the matching performance.

References

1. O'Reilly T, (2005) What is Web 2.0,
In http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
2. Cattuto C, Loreto V, Petronero L, (2007) Semiotic dynamics and collaborative tagging. In Procceedings of the National Academy of Sciences, 104:14611464
3. Halpin H, Shepard H, (1990) Evolving ontologies from folksonomies: Tagging as a complex system. In Complex Systems Summer School Project, http://www.ibiblio.org/hhalpin/homepage/ notes/ taggingcss.html.
4. Steels L, (2006) Semiotic dynamics for embodied agents. IEEE Intelligent Systems, 21:3238
5. Smeulders A W M, Worring M, Santini S, Gupta A, Jain R, (2000) Content-Based Image Retrieval at the End of the Early Years, In IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, number 12, pp. 1349–1380
6. National Information Standards Organization, (2004) Understanding Metadata. NISO Press, pp. 1-20
7. Wilensky R, (2000) Digital Library Resources as a Basis for Collaborative Work. Journal of The American Society for Information Science and Technology, 51(3):228245
8. Hobson P, Kompatsiaris Y, (2006) Advances in semantic multimedia analysis for personalised content access. Special Session on Advances in Semantic Multimedia Analysis for Personalised Content Access, IEEE International Symposium on Circuits and Systems
9. Golder S, Huberman A, (2006) The Structure of Collaborative Tagging Systems. Journal of Information Science
10. Begelman G, Keller Ph, Smadja F, (2006) Automated Tag Clustering: Improving search and exploration in the tag space. In Procceedings of Collaborative Web Tagging Workshop at the 15th WWW Conference, Edinburgh, Scotland

11. Grahl M, Hotho A, Stumme G, (2007) Conceptual Clustering of Social Bookmarking Sites. 7th International Conference on Knowledge Management, 356-364, KnowCenter,Graz, Austria.
12. Jaschke R, Hotho A, Schmitz Ch, Ganter B, Stumme G, (2006). TRIAS - An Algorithm for Mining Iceberg Tri-Lattices. In Proceedings of the 6th IEEE International Conference on Data Mining, 907-911
13. Gruber T, (2005) Folksonomy of Ontology: A Mash-up of Apples and Oranges. First On-Line conference on Metadata and Semantics Research MTSR
14. Knerr T, (2006) Tagging Ontology- Towards a Common Ontology for Folksonomies. Available at: http://code.google.com/p/tagont/
15. Newman R, (2005) Tag ontology design. Available at: http://www.holygoat.co.uk/projects/tags/
16. Brickley D, Miles A,(2005) SKOS Core Vocabulary Specification,W3CWorking Draft2. Available at: http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102
17. Schmitz P, (2006) Inducing Ontology from Flickr Tags. In Proceedings of the Collaborative Web Tagging Workshop at the 15th WWW Conference, Edinburgh, Scotland
18. Mika P, (2005) Ontologies are Us: A Unified Model of Social Networks and Semantics. In Proceedings of the 4th International Semantic Web Conference
19. Schmitz C, Hotho A, Jaschke R, Stumme G, (2006) Mining Association Rules in Folksonomies. In Proceedings of the (IFCS 2006), pages 261-270, Ljubljana
20. Specia L, Motta E,(2007) Integrating Folksonomies with the Semantic Web. In Proceedings of the 4th European Semantic Web Conference
21. Wu X, Zhang L, Yu Y, (2006) Exploring Social Annotations for the Semantic Web. In Proceedings of the 15th WWW Conference (WWW 2006), Edinnburgh, Scotland
22. 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
23. Michael S. Lew and Nicu Sebe and Chabane Djeraba Lifl and Ramesh Jain (2006) Content-based Multimedia Information Retrieval: State of the Art and Challenges, ACM Transactions on Multimedia Computing, Communications, and Applications, 2(1): 1-19
24. Pereira F. and Koenen R. (2001) MPEG-7: A standard for multimedia content description, Int. J. Image Graph, 1, 3, 527546
25. Lew M.S. (2001) Principles of Visual Information Retrieval, Springer, London, UK
26. Gevers T. (2001) Color-based retrieval. In Principles of Visual Information Retrieval, M. S. Lew, Ed. Springer-Verlag, London, UK, 1149
27. Ojala T., Pietikainen M. and Hardwood D. (1996) Comparative study of texture measures with classification based on feature distributionsm, Patt. Recogn. 29, 1, 5159
28. Jafari-Khouzani K. and Soltanian-Zadeh H. (2005) Radon transform orientation estimation for rotation invariant texture analysis, IEEE Trans. Patt. Analy. Machine Intell. 27, 6, 10041008
29. Bartolini I., Ciaccia P. and Patella M. (2005) WARP: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance, IEEE Trans. Patt. Analy. Machine Intellig. 27, 1, 142147
30. Srivastava A., Joshi S.H., Mio W. and Liu X. (2005) Statistical shape analysis: Clustering, learning, and testing, IEEE Trans. Patt. Analy. Mach. Intell. 27, 4, 590602
31. Sebastian T.B., Klein P.N. and Kimia B.B. (2004) Recognition of shapes by editing their shock graphs, IEEE Trans. Patt. Analy. Machine Intell. 26, 5, 550571
32. Vretos N., Solachidis V. and Pitas I. (2005) An MPEG-7 Based Description Scheme for Video Analysis Using Anthropocentric Video Content Descriptors, LECTURE NOTES IN COMPUTER SCIENCE, 3746, 725, Springer
33. Sebe N., Lew M.S. and Huijsmans D.P. (2000) Toward improved ranking metrics, IEEE Trans. Patt. Analy. Mach. Intell. 22, 10, 11321143
34. Jacobs D.W., Weinshall D. and Gdalyahu Y. (2000) Classification with nonmetric distances: Image hetrieval and class representation, IEEE Trans. Patt. Analy. Machine Intell. 22, 6, 583600

35. Beretti S., Del Bimbo A. and Vicario E. (2001) Efficient matching and indexing of graph models in content-based retrieval, IEEE Trans. Patt. Analy. Machine Intellig. 23, 10, 10891105
36. Cooper M., Foote J., Girgensohn A. and Wilcox L. (2005) Temporal event clustering for digital photo collections, ACMTrans. Multimedia Comput. Comm. Applica. 1, 3, 269288
37. Lindeberg T. (1998) Feature detection with automatic scale selection, Int. J. Comput. Vision, 30, 2, 79116
38. Lowe D. (2004) Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60, 2, 91110
39. Smeaton A. F., Over P. and Kraaij W. (2006) "Evaluation campaigns and TRECVid", In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006), MIR '06, ACM Press, New York, NY, 321-330
40. Maillot N., Thonnat M. and Boucher A. (2004) Towards ontology-based cognitive vision, Mach. Vis. Appl., 16, 1, 33-40
41. Hunter J., Drennan J. and Little S (2004) Realizing the Hydrogen Economy through Semantic Web Technologies, IEEE Intelligent Systems, 19, 1, Jan.-Feb., 40-47
42. Dasiopoulou S., Heinecke J., Saathoff C. and Strintzis M.G. (2007) Multimedia Reasoning with Natural Language Support, 1st IEEE International Conference on Semantic Computing (ICSC), Irvine, CA, USA
43. Aurnhammer M, Hanappe P, Steels L, (2006) Augmenting navigation for collaborative tagging with emergent semantics. In Proceedings of the 5th International Semantic Web Conference
44. Alvarado P, Doerfler P, Wickel J, (2001) Axon2 a visual object recognition system for non-rigid objects. In Proceedings of the International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)
45. Giannakidou E, Kompatsiaris I, Vakali A, (2008) SEMSOC: Semantics Mining on Multimedia Social Data Sources. In Proceedings of the 2nd IEEE International Conference on Semantic Computing, Santa Clara, CA, USA
46. Ghosh H, . Poornachander P, Mallik A, Chaudhury S, (2007) Learning ontology for personalized video retrieval. In International Multimedia Conference, Workshop on multimedia information retrieval on The many faces of multimedia semantics, Augsburg, Bavaria, Germany
47. Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T, (2007) How flickr helps us make sense of the world: context and content in community-contributed media collections. In In Proceedings of the 15th international Conference on Multimedia, Augsburg, Germany
48. Quack T, Leibe B, Van Gool L, (2008) World-scale mining of objects and events from community photo collections. In Proceedings of the 2008 international Conference on Content-Based Image and Video Retrieval, Niagara Falls, Canada
49. Crandall D, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the World's Photos. In Proceedings of the World Wide Web Conference, Madrid, Spain
50. Kennedy L, Naaman M, (2009) Less Talk, More Rock: Automated Organization of Community-Contributed Collections of Concert Videos. In Proceedings of the World Wide Web Conference, Madrid, Spain
51. Olivares X, Ciaramita M, van Zwol R,(2008) Boosting image retrieval through aggregating search results based on visual annotations. In Proceeding of the 16th ACM international conference on Multimedia, Vancouver, British Columbia, Canada
52. Lindstaedt S, Pammer V, Morzinger R, Kern R, Mulllner H, Wagner C, (2008) Recommending tags for pictures based on text, visual content and user context. In Proceedings of the Third International Conference on Internet and Web Applications and Services, Athens, Greece
53. Sigurbjornsson B, van Zwol R, (2008) Flickr tag recommendation based on collective knowledge. In Proceeding of the 17th international conference on World Wide Web, Beijing, China
54. Bumgardner J,(2006) Experimental colr pickr. Available at: http://www.krazydad.com/colrpickr/
55. Langreiter C, (2006) Retrievr. Available at: http://labs.systemone.at/retrie-vr/

56. Maguitman A, Lord P.W, Menczer F, Roinestad H, Vespignani A, (2005) Algorithmic Detection of Semantic Similarity. In Proccedings of the 14th international conference on World Wide Web , (WWW'05),pages 107-116
57. Wu Z, Palmer M, (1994) Verm semantics and lexical selection. In Proceedings of the 32nd annual meeting of the association for computational linguistics, pages = 133-138. New Mexiko, USA.
58. Martnez J.M, "Overview of the MPEG-7 Standard (v4.0)", ISO/MPEG N3752
59. B. S. Manjunath, Philippe Salembier, Thomas Sikora (2002) Introduction to MPEG-7: Multimedia Content Description Interface, John Wiley & Sons, Inc. New York
60. MPEG-7 Visual Experimentation Model (XM), Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4062, Mar., 2001.
61. Fellbaum C, (1990) WordNet, an electronic lexical database. The MIT Press
62. Chatzilari E, Nikolopoulos S, Giannakidou E, Vakali A, Kompatsiaris I. (2009) Leveraging Social Media For Training Object Detectors. In Proceedings of the 16th International Conference on Digital Signal Processing, Special Session on Social Media, Santorini, Greece
63. Larsen B. and Aone C. (1999) Fast and effective: Text mining using linear-time document clustering, Proc. of 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, (KDD99), pages 1622, August
64. Xu R.(2005) Survey of Clustering Algorithms. In IEEE Transactions on Neural Networks, Vol.16, No.3, May
65. Buturovic Adis (2005) MPEG 7 Color Structure Descriptor for visual information retrieval project VizIR1. Institute for Software Technology and Interactive Systems, Technical University Vienna
66. B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, and Akio Yamada (2001) Color and Texture Descriptors, IEEE Trans. On Circuits and Systemsfor Video Technology, vol. 11, No. 6