Combining multi-modal features for social media analysis

Spiros Nikolopoulos, Eirini Giannakidou, Ioannis Kompatsiaris, Ioannis Patras, Athena Vakali

Abstract In this chapter we discuss methods for efficiently modeling the diverse information carried by social media. The problem is viewed as a multi-modal analysis process where specialized techniques are used to overcome the obstacles arising from the heterogeneity of data. Focusing at the optimal combination of low-level features (i.e., early fusion), we present a bio-inspired algorithm for feature selection that weights the features based on their appropriateness to represent a resource. Under the same objective of optimal feature combination we also examine the use of pLSA-based aspect models, as the means to define a latent semantic space where heterogeneous types of information can be effectively combined. Tagged images taken from social sites have been used in the characteristic scenarios of image clustering and retrieval, to demonstrate the benefits of multi-modal analysis in social media.

Spiros Nikolopoulos
Informatics & Telematics Institute, Thermi, Thessaloniki, Greece and School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, UK, e-mail: nikolopo@iti.gr

Eirini Giannakidou
Informatics & Telematics Institute, Thermi, Thessaloniki, Greece and Department of Computer Science, Aristotle University of Thessaloniki, Greece, e-mail: igiannak@iti.gr

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

Ioannis Patras
School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, UK Tel. +44 20 7882 7523, Fax: +44 20 7882 7997, e-mail: i.patras@eecs.qmul.ac.uk

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

1 Introduction

Content sharing through the Internet has become a common practice for the vast majority of web users. Due to the rapidly growing new communication technologies, a large number of people all over the planet can now share, tag, like or suggest a tremendous volume of multimedia content, that we generally refer to as social media. One of the most outstanding features for social media is their intrinsic multi-modal nature that opens-up new opportunities for content analysis and consumption. The goal of this chapter is to discuss the use of mutli-modal approaches for social media analysis and demonstrate their effectiveness in certain application scenarios.

1.1 Social Media

The recent advances of Web technologies have effectively turned ordinary people into active members of the Web. Web users act as co-developers and their actions and collaborations with one another have added a new social dimension on Web data. Social media are commonly described by a high diversity of features. For instance, an image in Flickr is associated with the tags that have been assigned to it, the users that seem to like it and mark it as favorite, the visual features that describe the visual content of the image, and possibly spatial or temporal information that denote the spatial and temporal context of this particular image. Even though all these facets of information are not combined naturally with each other, still they carry knowledge about the resource, with each facet providing a representation of the particular resource in a different feature space.

This information provides added value to the shared content and enables the accomplishment of tasks that are not possible otherwise. Exploiting the information carried by social media can help tackle a variety of issues in different disciplines, such as content consumption (e.g., poor recall and precision), knowledge management (e.g., obsolescence, expertise), etc. However, the exploitation of such information is a big departure from traditional methods for multimedia analysis, since managing the diversity of features poses new requirements. For example, semantic analysis has to fuse information coming both from the content itself, the social context and the emergent social dynamics. Moreover, the unconstrained nature of content uploading and sharing has resulted in large amounts of spammy and noisy content and metadata, thus considerably compromising the quality of data to be analyzed. All these have motivated an increasing interest on investigating methods that will be able to exploit the intrinsic multi-modality of social media.

1.2 Multi-modal analysis

Semantic multimedia indexing has been recognized as a particularly valuable task for various applications of content consumption. Current literature has made considerable progress in this direction especially for uni-modal scenarios. However, it is generally accepted that multi-modal analysis has the potential to further improve this process, provided that the obstacles arising from the heterogeneous nature of

different modalities can be overcome. This is based on the fact that independently of whether these pieces of information act cumulatively or complementary, when combined; they encompass a higher amount of features that can be exploited to improve the efficiency of the performed task. Depending on the level of abstraction where the combination of different modalities takes place, we can distinguish between the result level (or late fusion) and the feature level (or early fusion) approaches. In the result-level approaches, features from each data source are initially extracted separately and, still separately, transformed into conceptual information to be utilized by the fusion process. In the feature-level approaches, features are extracted separately from each modality but they are integrated into a jointed, unique representation, before the employed approach transform them into conceptual information and deliver them to the fusion process. In this chapter we discuss methods that belong to the feature level of abstraction.

In feature level approaches the need to obtain a jointed, unique representation for the multimedia object demands for techniques that will manage to handle the very different characteristics exhibited by the different types of data. This is true both in terms of the nature of raw features (e.g., sparse, high-dimensional word co-occurrence vectors extracted from text descriptions, compared to usually dense and low-dimensional descriptors extracted from visual content), as well as in terms of their semantic capacity (e.g., while abstract concepts like “freedom” are more easily described with text, concrete concepts like “sun” are more easily grounded using visual information). Based on the above one can pursuit a solution to the multi-modal analysis problem by following two different approaches, that although similar in scope they differ in the way of handling the features. The first case, usually referred by the name “feature selection”, aims at assigning appropriate weights to the features of the initial space. To do so, the correlations among objects in each individual feature space are examined. Although the dimensionality of the resulting feature space is identical with the dimensionality of the initial feature space, the process results in groups of related multimedia objects which lie in sub-variations of the initial feature space. The second case, that can be named as “feature combination”, aims at defining a new feature space where the projection of the initial features will yield an improved and homogeneous representation of the multimedia object. In this case the dimensionality of the resulting feature space is typically smaller than the dimensionality of the initial space. In this respect we can claim that “feature selection” approaches primarily target at exploiting the different levels of semantic capacity exhibited by the different feature spaces, while the “feature combination” approaches aim at removing the problems arising from the heterogeneity of data and benefit from their efficient combination.

1.3 Examined approaches & applications

In this chapter we discuss methods falling under both feature handling approaches and present in detail one indicative method for every case. In the first case, we examine an ant-inspired algorithm for feature selection which is based on Ant Colony Optimization (ACO) metaheuristic. This algorithm is used to perform clustering on tagged images and its goal is to define appropriate weights so that each feature space weight sufficiently captures the local correlation of the resources along the specific

dimension. This is done by simulating the way the ant colony finds the shortest path for solving a clustering problem in multi feature spaces. The algorithm was applied on a dataset of Flickr images in which two feature spaces are taken into consideration: i) tag features, and ii) visual features. Our experimental study has shown that the ant-based clustering algorithm performs better than a typical clustering algorithm that relies on predefined feature spaces.

For the case of feature combination we examine an algorithm based on aspect models. These models assume that the content depicted in an image can be expressed as a mixture of multiple latent topics, defining a new, "semantically enhanced" feature space. Thus, the meaningful combination of visual and tag information is performed by projecting the initial features to the space of latent topics. In addition, in order to further exploit the cross-modal relations existing between the tag and visual content referring to the same abstract meaning, we examine the employment of a hierarchical version of the aspect model. Our experiments on using this algorithm to retrieval relevant images on a dataset of Flickr images, have demonstrated the effectiveness of the aspect models to provide an appropriate space for the combination of heterogeneous features.

Both clustering and retrieval scenarios that have been used to demonstrate the effectiveness of the examines approaches are particularly important in the context of social media. Indeed, the multi-faceted organization of content as well as the ability to retrieve relevant resources, act as the basis for various social network related applications. For instance, tag recommendation for images and friend recommendation for users are applications that both need to consider the available content from many different aspects. Multi-modal analysis is ideal for this task since different types of information can be combined and exploited. Additionally, many of the functionalities provided by social networks are essentially build on top of the outcome of a clustering or a relevance ranking task.

The remaining of this chapter is organized as follows. Section 2 reviews the related literature on multi-modal analysis. Section 3 provides a formulation for the feature selection problem and presents how an ant-inspired algorithm can be used to solve the formulated optimization task. Section 4 discusses the use of aspect models for achieving optimal feature combination and shows how the employment of a topic hierarchy manages to outperform all other combination schemes. Conclusions and avenues for future research are presented in Section 5.

2 Related Work

In this section we review related works from the field of multi-modal analysis and social media. Initially we present some methods that can be considered as typical examples of multi-modal analysis. Subsequently, we review the related literature of multi-modal analysis methods that have been mainly proposed to facilitate the analysis of social media. Finally, we refer to a set of works that emphasize on the range of social media applications that can benefit from multi-modal analysis.

2.1 Methods for multi-modal analysis

Under the general objective of optimally combining heterogeneous types of information, current literature proposes methods that tackle the problem by exploiting the statistical properties of the data. Different types of transformations have been proposed for determining a space where features extracted from heterogeneous sources can be optimally combined. The work presented by Mogalhães and Rüger [30] is an indicative example of this category where information theory and a maximum entropy model are utilized to integrate heterogeneous data into a unique feature space. The authors use a minimum description length criterion to find the optimal feature space for text and visual data. Then, both feature spaces are merged in order to obtain a unique continuous feature space. Evaluation is performed on a document retrieval scenario based on only text, only visual and combined data. In [40] Wu et al. work on the same direction by introducing a method that initially finds statistical independent modalities from raw features and subsequently applies super-kernel fusion to determine their optimal combination. More specifically, an independent modality analysis scheme is initially presented, which applies Independent Component Analysis (ICA) on the raw set of features. Then, the modalities are fused by employing a scheme that finds the best feature combination through supervised learning. In [26] Li et al. present several cross-modal association approaches under the linear correlation model: the latent semantic indexing (LSI), canonical correlation analysis (CCA) and Cross-modal Factor Analysis (CFA). They claim that the proposed CFA approach manages to identify and exploit the intrinsic associations between different modalities by treating features from different modalities as two distinguished subsets and focusing only on the semantic patterns between these two subsets.

Another direction to facilitate mining of objects described in many modalities is to distinguish between important and less important features in the object's description. The techniques that detect clusters of objects in all possible variations of subspaces fall in the category of Subspace clustering [3]. Such techniques extract cluster structures that are hidden by noisy dimensions and aim at separating relevant and irrelevant dimensions locally (in each cluster). One of the first approaches to subspace clustering is CLIQUE (CLustering In QUEst) described in [2]. CLIQUE is a grid-based algorithm using an apriori-like method to recursively navigate through the set of possible subspaces in a bottom-up way. Then, a number of slight modifications of CLIQUE have been proposed like ENCLUS- ENtropy-based CLUStering, differing mostly at the criterion used for the subspace selection is the algorithm ([9]).

A complete review of subspace clustering techniques for high-dimensional data can be found in [33].

2.2 Multi-modal analysis of social media

The multi-modal aspect that is intrinsic in social media prompt many researchers to propose specialized methods for their multi-modal analysis. In this category of works we classify the ones relying on the use of aspect or topic models [20] and the definition of a latent semantic space. In [27] the authors use a model based on Probabilistic Latent Semantic Analysis (pLSA) [21] to support multimodal image retrieval in Flickr, using both visual content and tags. They propose to extent the standard single-layer pLSA model to multiple layers by introducing not just a single layer of topics, but a hierarchy of topics. In this way they manage to effectively and efficiently combine the heterogeneous information carried by the different modalities of an image. In a similar fashion the authors of [38] propose an approach for multimodal characterization of social media by combining text features (e.g., tags) with spatial knowledge (e.g., geotags). The proposed approach is based on multimodal Bayesian models which allow to integrate spatial semantics of social media in well-formed, probabilistic manner. As in the previous case the authors aim to explain the observed properties of social media resources (i.e., tags and coordinates) by means of a Bayesian model with T latent topics. The approach is evaluated in the context of characteristic scenarios such as tag recommendation, content classification, and clustering.

Only recently, there has been an increasing interest on extending the aspect models to higher order through the use of Tensors [24]. Under this line of works we can mention the tag recommendation system presented in [39] that proposes a unified framework to model the three types of entities that exist in a social tagging system: users, items and tags. These data are represented by a 3-order tensor, on which latent semantic analysis and dimensionality reduction is performed using the Higher Order Singular Value Decomposition (HOSVD) technique [25]. Consequently, tags are recommended based on the proximity of the resources in the latent semantic space. A 3-order tensor is also used by the authors of [16] that propose an approach to capture the latent semantics of Web data by means of statistical methods. In order to do that the authors apply the PARAFAC decomposition [19] which can be considered as a multi-dimensional correspondent to a singular value decomposition of a matrix. In this case the extracted latent semantics are used for the task of relevance ranking and producing fine-grained and rich descriptions of Web data.

Furthermore, the problem of multi-modality has been given rise to applying subspace clustering techniques in the social media environments. Currently such approaches are rather few in this area. More specifically, in [5] the authors use *cluster ensemble* approaches [12] to get clusters of social media that are associated with events. They form a variety of representations of social media resources, using different context dimensions and combine these dimensions into a single clustering solution. Another novel subspace clustering technique that is based on ACO metaheuristic is presented in [34]. The method performs the subspace clustering on the visual features that describe an image and is tested in a Flickr dataset.

2.3 Social media applications with multi-modal analysis

The multi-modal analysis has been used as the core analysis component of various different applications in social media. First of all, the fact that most people are fond of uploading and sharing content in social media environments motivates research efforts towards better browsing and retrieval functionalities in such environments. Furthermore, the intrinsic limitations of these systems (e.g. tag ambiguities, erroneous metadata or lack of metadata, etc.) addresses the need for exploitation of features in multi modalities, in order to provide users with as much possible relevant content. To this end, in [4], the authors present a method for a social image database browsing and retrieval by exploiting both tag and visual features of the images 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. Another approach in this direction is met at [17], where the authors present clustering algorithms that improve the retrieval in social media by exploiting features from multi spaces. Specifically, a two-step clustering approach is proposed that uses tag features at the 1st step, in order to get resources clusters that refer to certain topics, and visual features at the 2nd step, in order to further "clean" and improve the precision to the extracted clusters. The use of both visual and text features is also described in [32], where the authors deploy the visual annotations, also known as "notes" in Flickr, and it is shown that the retrieval of social media content improves significantly by combining tags and visual analysis techniques.

A number of works have addressed the problem of identifying photos from social media environments that depict a certain object, location or event [23, 35, 11]. In [23] they analyze location and temporal features 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 clustering on their visual features, in order to capture distinct viewpoints of the object of interest. The same authors in [22] combine tags with content-based features and analysis techniques, in order to get groups of music events photos. Likewise, in [35, 11, 6] 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 the described methods that perform object or POI identification can be used for training of multimedia algorithms, whereas these methods that extract social media content associated with particular events can be exploited for faceted browsing of events and related activities in browsers.

Most of the aforementioned methodologies can be exploited for tag recommendations in the sense that they extract tags associated to a particular event, object, location or, in general, cluster of related resources. The problem of tag recommendation has been further studied in [28], 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 [36], where the authors suggest four methods for ranking candidate tags and in addition, they present the semantics of tags in Flickr.

3 Combining heterogeneous information using Ant Colony Optimization

In this section, we are going to tackle the problem of multi-modality in social media, using a feature selection technique. The method we present falls in the category of the so-called subspace clustering approaches that identify clusters of high-dimensional objects and their respective feature subspaces. Specifically, we aim at providing a clustering solution that assigns a feature weight to each dimension on each cluster, based on the correlation of the cluster's objects along the specific dimension. As we will show in 3.1, this is a combinatorial optimization problem and we approximate it using the Ant Colony Optimization (ACO) metaheuristic [13].

Ant colony optimization has been applied successfully to a large number of difficult discrete optimization problems including the traveling salesman problem, the quadratic assignment problem, scheduling, vehicle routing, etc., as well as to routing in telecommunication networks ([8, 14, 7]. Although data clustering techniques for social media have been heavily researched, little research has been dedicated on the use of bio-inspired algorithms for extracting information from this type of content. In this section, we employ an ACO algorithm and demonstrate how it can be used to tackle the problem of combining multi-feature spaces in the clustering of social media problem. We tested the approach on a Flickr dataset and our preliminary experiments show promising results with respect to other baseline approaches.

3.1 Problem Formulation

Social media are commonly described by a high diversity of features. To benefit from all this available information, we assume, here, that each resource r is represented by D different feature vectors:

$$r = (F_1, F_2, \ldots, F_D)$$

where F_i, $1 \leqslant i \leqslant D$ is a feature vector from a corresponding feature space $\mathfrak{F}_i$. We can now define the distance between two resources by considering appropriate distance measures for each feature space. For instance, we calculate distances in the tag space, based on tag co-occurrence, whereas we use Euclidean distance to capture the difference in the geographical coordinates between two resources. Thus, given D valid distance measures between the corresponding D feature vectors of the resources r_1 and r_2, we can get their distance as:

$$d^w(r_1, r_2) = \sum_{i=1}^{D} w_i d_i(F_{i_1}, F_{i_2})$$

where d_i is the distance measure employed in feature space $\mathfrak{F}_i$, $1 \leqslant i \leqslant D$ and w_i is a feature weight that determines the influence of the resources' $i-th$ feature vector to the calculation of the overall distance. In other words, the use of feature weights allows to be given different degree of gravity along each dimension. It holds $\sum_{i=1}^{D} w_i = 1$ and $w_i > 0$.

An example scenario that shows the importance to detect clusters in different variations of feature space in social media follows. Assume that there are 3 groups of social content: Group A that contains photos depicting waves, Group B with photos depicting people waving, and Group C depicting various places from Paris. The tags that describe each group are *wave*, *wave*, *paris*, respectively[1]. If we try to detect clusters using only tag features, we will get a cluster that contains $A+B$ together and a separate cluster C. On the other hand, if we use only visual features, the cluster C will not be obtained. Using both modalities as equally important misses also Group C, and, in general, this approach fails to detect clusters that are associated with abstract concepts, as we will show in our Experimental Study in 3.3. We claim that managing to define appropriate feature weights for each feature space is a way to obtain in separate clusters the content that is contained in each of the groups A, B, and C.

In this section, the purpose is to perform clustering on social media resources by optimally combining multi-feature information. They key idea is not to combine all the features together, but to examine local correlations between resources across each dimension and, thus, detect resources' clusters in all feature subspaces. Such techniques are known as subspace clustering and the resulting clusters may refer to different feature subspaces. More formally, we aim at providing solution to the following problem.

Problem 1. Given a set of N social media resources described by features from D different spaces, a set of D distance measures $d_1, d_2, \ldots, d_D$, one for each feature space, and an integer K, find a $K-$partitioning of resources $C_1, C_2, \ldots, C_K$, such that $\sum_{i=1}^{K} \sum_{r_1, r_2 \in C_i} d^w(r_1, r_2)$, where d^w a weighted distance in each cluster, is minimized.■

In order to obtain the d^w, we should define appropriate values for feature weights w_i, $1 \leq i \leq D$, so that each feature weight sufficiently captures the local correlation of the resources along the specific dimension. To do so, we employ an ant-inspired algorithm, which is based on ACO metaheuristic. The method of ACO metaheuristic technique which was proposed by Dorigo is a model of the ant behavior, which is used for combinatorial problems [14]. In the next section, we present a modification of this algorithm which can be employed to solve the social media clustering in multiple feature spaces problem.

3.2 The proposed framework

Ant algorithms were inspired by the behavior of real ants when searching for food. When an ant finds food, it releases a chemical substance called pheromone along the path from the food to the nest. Pheromone provides an indirect communication among the ants, since ants typically follow pheromone trails. The amount of pheromone that exists in each path is proportional to the number of ants that have used this path. Pheromone evaporates in time, causing trails and paths that are no longer followed by ants to extinguish.

[1] As many users find tedious the tagging process, the scenario that most photos in each group have been assigned only one tag is not far from reality

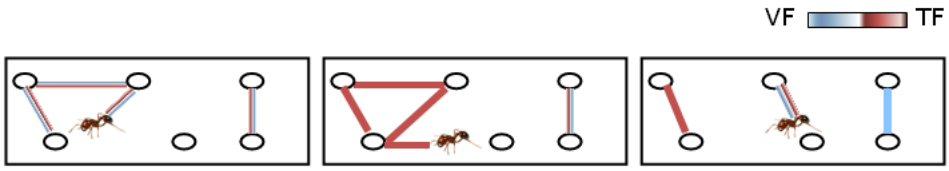


Fig. 1 Combining multi-features in social media clustering using ACO.

This pheromone-driven communication between ants have been modeled to solve a number of research problems, one of the most well-known being the Traveling Salesman Problem (TSP). The fact that pheromone evaporates in time causes pheromone trails in longer paths to weaken, as it takes more time for the ants to cross them. On the contrary, a short path is traversed faster, and, thus, the pheromone trail on this path becomes stronger, as it is laid on the path as fast as it can evaporate and many ants follow the trail and reinforce it. This behavior in ant colonies can be used to find shortest paths between nodes in a graph and, thus, providing good solutions to TSP in the following way: Agents are modeled as ants who cross the graph, as they search for food. Initially, the ants are moving randomly. If they meet pheromone trails, they follow them until they visit all the nodes in the graph, constructing, this way, each ant an incremental solution to the problem. When an ant has visited all the nodes, it releases pheromone inversely proportional to the length of the path. Thus, the shorter the path the bigger amount of pheromone is released, attracting other ants to follow the particular path. It is important to note that pheromone evaporation prevents sticking to local minima and allows a dynamic adaptation when the problem changes.

In this section, we present the way the ant colony behavior finding the shortest path can be simulated, in order to solve a clustering problem in multi-feature spaces. The idea is roughly depicted in Figure 1, in which two feature spaces are taken into consideration: i) tag features, and ii) visual features of resources and the description follows: A large number of virtual ants are sent to explore many possible clustering solutions of social media resources. The resources are depicted as graph nodes and the ants join two nodes with an edge if they decide to assign them in the same cluster. The color of the edge shows the weight given by each ant to each feature space individually. The weight is based on the pheromone that there is in each edge. Initially, both feature spaces are given equal weight. Each ant probabilistically assigns each resource to a cluster, based on a measure combining the distance to the cluster center in each feature space individually and the amount of virtual pheromone deposited on the edge to the resource. The ants explore, depositing pheromone on each edge that they cross, until they have all completed their clustering. At this point, the pheromone to each clustering solution is updated (global pheromone updating), so that the edges that have been crossed by many ants become bolder, whereas the remaining ones (that haven't been selected by many ants) become thinner. The amount of pheromone deposited is proportional to the resources correlation along each feature space: the bigger the correlation in tag/visual space, the more pheromone is deposited. The color shows the correlation in each feature space, that is: red color denotes more weight to tag features, whereas blue-colored paths signify clusters that contain objects with high visual similarity.

A pseudocode description of the approach is presented next. At first, an initialization procedure takes place, during which: i) each ant initializes K centroids randomly. Each centroid c_i is selected to be represented as $(c_{i1}, c_{i2}, \ldots, c_{iD})$, where c_{ij} is a vector from the feature space $\mathfrak{F}_j$, $1 \leq i \leq K$, and $1 \leq j \leq D$ *(lines 7-10)*, ii) the pheromone amount of all graph edges are set to 1 *(line 3)*, iii) the feature weights w_i are set to 0.5, which equals to $1/D$ *(lines 4-6)*, iv) the parameters ρ:pheromone evaporation factor, h:constant used to determine the influence of the distance measure against the pheromone value in the cluster assignment process, are initialized *(line 2)*.

Then, the clustering process begins during which each ant will decide what edges to cross in the graph and what color to paint them. To do so, the following process is repeated for all resources: i) each ant calculates the distance to each cluster centroid in each feature space individually *(lines 15-18)*, ii) considering the feature weights that are already calculated from the previous iteration of the algorithm, each ant estimates an overall distance from each resource to each cluster centroid *(line 19)*, iii) given the overall distance $d^w(r,c)$ calculated in the previous step, the pheromone amount that there is currently on graph edges and the constant h, each ant determines the probability that a resource r should be assigned in a cluster with centroid c, as follows:

$$p(r,c) = \frac{\tau(r,c) \cdot h/d^w(r,c)}{\sum_{i=1}^{K} \frac{\tau(r,c_i) \cdot h}{d^w(r,c_i)}}$$

The ant assigns the resource to the cluster with the highest probability *(line 20)*. This process is illustrated in the graph in Figure 3.2, as an ant marking an edge between a resource and the other resources already in the cluster. The color of the edge depends on the values of the feature weights w_i.

Having performed the clustering process, new feature weights are calculated for each cluster with centroid c, based on the correlations that there are among the resources in each feature space in the cluster, as follows:

$$w(c,i) = \frac{\sum_{r \in c} d_i(r,c)}{\sum_{r \in c, l=1}^{D} d_l(r,c)}$$

for $1 \leq i \leq D$ and $d_i(r,c)$ is the distance from the resource r to the cluster centroid c in the feature space $\mathfrak{F}_i$ *(line 23-25)*.

Next, new centroids are calculated, based on the assignments in each cluster *(line 26)*. After all ants have done their clustering, the pheromone amount to all solutions is recalculated *(lines 30-32)*. To do so, the quality of each solution needs to be estimated, so that ants that provided good solutions generate more pheromone. The measure we use for ranking the solutions is derived from the definition of clustering, as given in [41] according to which the resources that belong in one cluster should be closely similar to each other, according to some metric of similarity, while the ones that belong to different clusters should be dissimilar. Thus, the most "efficient" ant generates a clustering where: i) in each cluster with centroid c the intra-cluster distance is minimized , that is $IntraDistance_{ant} = \min \sum_{r \in c} \sum_{i=1}^{D} w(c,i) \cdot d_{\mathfrak{F}_i}$, and ii) the inter-cluster distance is maximized, that is: $InterDistance_{ant} = \max \sum_{r_1 \in c_1, r_2 \in c_2, c_1 \neq c_2} \sum_{i=1}^{D} d_{\mathfrak{F}_i}$. We assume that the quality of each solution is given by:

$$\mathbf{q}_{ant} = \frac{InterDistance_{ant}}{IntraDistance_{ant}}$$

Calculating these measures for each solution, we update the current pheromone of each path by considering the total number of ants that have used that path and the quality of their solution. That is:

$$\Delta\tau(r,c) = \sum_{ant} \mathbf{q}_{ant}, \forall ant : (r,c) = 1$$

where $1 \leq r \leq N$, $1 \leq c \leq K$. Furthermore, during the global pheromone update, the current pheromone evaporates at a constant rate ρ at each iteration step. The presented ant-based clustering process *(lines 14-32)* is repeated for $NumberOfIteration$ times until the Problem 1 is satisfied.

Algorithm 1 The ANT-BASED clustering algorithm combining features from multi-feature spaces.

Input: N number of social media resources, S number of ants, K number of clusters, T number of iterations, D number of feature spaces $\mathfrak{F}_1, \mathfrak{F}_2, \ldots, \mathfrak{F}_\mathfrak{D}$
Output: A set $C = \{C_1, \ldots, C_K\}$ of K resources subsets and a feature weight vector $(w_{\mathfrak{F}_1}, w_{\mathfrak{F}_2}, \ldots, w_{\mathfrak{F}_\mathfrak{D}})$that best describes each cluster.

1: */*Initialization*/*
2: initialize the parameters h, ρ
3: initialize the pheromone value τ to 1
4: **for** $\mathfrak{F}_\mathfrak{i}$ in $[\mathfrak{F}_1, \ldots, \mathfrak{F}_\mathfrak{D}]$ **do**
5: $w_i \leftarrow 1/D$ */*Initialize feature weights*/*
6: **end for**
7: **for** *ant* in $[1,S]$ **do**
8: **for** k in $[1,K]$ **do**
9: centroids $\leftarrow$ randomly $c = (c_1, c_2, \ldots, c_D)$, where $c_i \in \mathfrak{F}_\mathfrak{i}$ and $1 \leq i \leq D$
10: **end for**
11: **end for**
12: */*Ant-based Clustering*/*
13: **while** iter $<$ $NumberOfIterations$ **do**
14: **for** *ant* in $[1,S]$ and r in $[1,N]$ **do**
15: **for** c in $[1,K]$ **do**
16: **for** $\mathfrak{F}_\mathfrak{i}$ in $[\mathfrak{F}_1, \ldots, \mathfrak{F}_\mathfrak{D}]$ **do**
17: $d_{\mathfrak{F}_\mathfrak{i}}(r,s) \leftarrow CalculateDistance$ in $\mathfrak{F}_\mathfrak{i}$
18: **end for**
19: $d^w(r,c) \leftarrow CalculateOverallDistance(d_{\mathfrak{F}_\mathfrak{i}}, w_i)$
20: *p(r, c)* $\leftarrow CalculateProbability(\tau, d^w, h)$
21: **end for**
22: **for** c in $[1,K]$ **do**
23: **for** $\mathfrak{F}_\mathfrak{i}$ in $[\mathfrak{F}_1, \ldots, \mathfrak{F}_\mathfrak{D}]$ **do**
24: $w_(c, \mathfrak{F}_\mathfrak{i}) \leftarrow CalculateClusterCorrelation$ in $\mathfrak{F}_\mathfrak{i}$
25: **end for**
26: centroids $\leftarrow CalculateNewCentroids$
27: **end for**
28: **end for**
29: */*Global pheromone update*/*
30: **for** r in $[1,N]$ and c in $[1,K]$ **do**
31: $\tau(r,c)_{iter} = \rho \cdot \tau_{(iter-1)}(r,c) + \sum_{ant=1}^{S} \tau_{iter}(r,c)$
32: **end for**
33: **end while**

3.3 Experimental Study

In order to test the described algorithm, a dataset from Flickr online photo management and sharing application was crawled. As we were interested, initially, to check the functionality of the algorithm, we conducted experiments on a rather small dataset and examine how the ant-based clustering algorithm performs better than a typical clustering algorithm that relies on predefined feature spaces. Thus, the dataset was restricted to 3000 images (size 500x735) that depict cityscape, seaside, mountain, roadside, landscape and sport-side locations[2].

At first, we applied typical clustering algorithms (K-Means, Hierarchical and Cobweb) for various values of K, based only on tag features. The distance measure used is a combination of tag co-occurence and WordNet distance [17]. We examine the extracted clusters manually and observed that many clusters had poor accuracy, i.e. they contained resources not depicted related themes, although sharing related tags. Figure 2 shows some indicative snapshots of this type of clustering. Specifically, the limitation of algorithms relying solely on tag information to handle ambiguous terms is shown. A snapshot of a cluster containing social media resources about *Paris* is shown in (b).

(a) Cluster around the topic *rock* (Ambiguous tag)

(b) Cluster around the topic *Paris*

Fig. 2 Indicative outcome of clustering using tag features.

To minimize the intrinsic shortcomings of tagging features, we embedded in the clustering process the visual features of the images. More specifically, we performed clustering using both tag and visual features in a sequential way: first we apply tag-based clustering and then clustering based on the visual features of the resources [17]. The visual feature extraction was based on the MPEG-7 standard [31] 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 was utilized in our approach [1]. The descriptors used were the Color Structure Histogram *(CSH)* and Edge Histogram *(EH)* descriptors, chosen due to their effectiveness in similarity re-

[2] For Flickr resources and metadata download the Flickr API along with the utility wget were used.

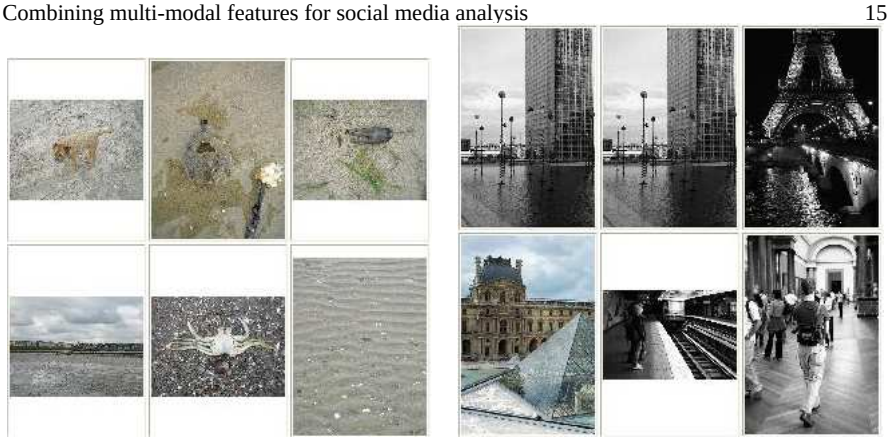
trieval. Their extraction was performed according to the guidelines provided by the MPEG-7 XM and then, an image feature vector was produced, for every resource, by encompassing the extracted MPEG-7 descriptors in a single vector. Thus, typical clustering algorithms could be applied, using the distance functions that are defined in MPEG-7 XM [17]. A set of users checked the extracted clusters manually and assessed their quality. The results showed that in many cases the accuracy was improved. This was especially true for clusters that contained ambiguous terms (e.g. rock - stone, rock -music). Especially for cases that the two senses of the ambiguous tag differed a lot visually, the algorithm succeeded to distinguish the different senses of the ambiguous tag, by dividing the corresponding resources into different clusters(cf. Figure 3). However, there were cases that the combination of tag and visual features worsen the clustering outcome. An example of such case is the *Paris* cluster and, generally, clusters whose topic is not related to a particular visual representation (abstract concepts). Indeed, we saw that a cluster describing *Paris* was extracted based on tag features of resources. This cluster no longer exists, if we employ the visual features to the clustering. As shown in Figure 3, this type of clustering succeeds in assigning resources with uniform visual appearance together. This way though it misses clusters of abstract resources that refer to the same topic but differ visually.

(a) Cluster around the topic *rock* (sense:stone)

(b) Cluster around the topic *rock* (sense:music)

Fig. 3 Indicative outcome of clustering using tag features and visual features.

The aforementioned example shows that the equal consideration of diverse features describing a resource does not always yield the optimal results. Some cluster resources can be only extracted by using a specified combination of features. We apply the presented ACO-based method using two modalities of the images, i.e. textual and visual. For the representation of images in the tag space we used the approach described in [18] and for their representation in a visual space we used their Color Structure Histogram *(CSH)* and Edge Histogram *(EH)* descriptors, as described above. We conducted a number of experiments with different values of the parameters ρ and h. Figure 4 shows indicative clustering results of the ant-inspired algorithm. It can be seen that the algorithm sufficiently captures clusters in different feature subspaces and it managed to handle the *Paris* cluster well.



(a) Cluster around the topic *sea*
$w_{TF} = 0.136, w_{VF} = 0.86, \rho = 0.85, h = 0.75$

(b) Cluster around the topic *Paris*
$w_{TF} = 0.84, w_{VF} = 0.002, \rho = 0.85, h = 0.75$

Fig. 4 Indicative outcome of clustering using presented ACO algorithm.

The presented experiments are preliminary and were performed to evaluate qualitatively the use of ACO methods in social media clustering. We reached useful conclusion that sum up as follows. It is apparent that the restriction in one feature space deprives information that can come out handy in the task of social media clustering. On the other hand, by considering all the feature spaces equally important we may miss clusters of related object that have similarity using a specific combination of features. In this section we proposed applying subspace clustering for obtaining clusters of social media in various feature subspaces. An ant-inspired algorithm was presented to realize this task. Future work involves the testing of the algorithm on larger datasets.

4 Combining heterogeneous information using aspect models

As an alternative approach of feature selection, efficient multi-modal analysis can be also achieved by employing approaches for feature combination. In contrast to feature selection that handles multi-modal features by assigned them appropriate weights, feature combination aims at defining a new feature space suitable for projecting the multi-modal features. One very popular scenario for these approaches is the retrieval of relevant images using an example image. The multi-modal aspect of the image retrieval scenario has been primarily boosted by the rapid growth of social networks, which resulted in the abundant availability of tagged images on the Web. The existence of tags allow these images to be indexed based on both their visual and tag information, which is expected to better facilitate the retrieval of semantic relevant images. However, as will become apparent in the remaining of this section, the heterogeneous nature of visual content and tags makes their combination a challenging problem.

4.1 Motivation & approach

Most existing content-based image retrieval systems are based either on the visual or textual information contained in the query image to retrieve relevant images. It is common consensus that combining the information carried by both modalities should lead to the improvement of retrieval results. However, the straightforward combination of both modalities that subsumes the words (i.e., type of features detailed in Section 4.2 that are extracted from visual or textual content and resemble the use of normal words in our every day language) extracted indiscriminately from both modalities to be parts of one common word set (usually called vocabulary), does not lead to the expected improvements. One possible reason is that by indiscriminately placing features extracted from heterogeneous information sources into a common feature space, the obtained image representations are likely to be dominated by one of the combined modalities or loose connection with their semantic meaning. Moreover, by simply considering the extracted words as independent elements of the same large vocabulary, we automatically assume that all these words are mutually independent. In this way we disregard the fact that in most cases, the words derived from the image visual and tag information are essentially two different expressions of the same abstract meaning. This implies a type of dependence that needs to be taken into consideration when designing the image representation scheme.

In order to facilitate the combination of heterogeneous modalities and avoid the aforementioned problems, we need to determine a feature space that will allow the extracted words to be expressed as a mixture of meanings. For this purpose current literature proposes the use of Probabilistic Latent Semantic Analysis (pLSA)-based [21] aspect or topic models that allow to map a high-dimensional word distribution vector to a lower-dimensional topic vector (also called aspect vector). These models assume that the content depicted by every image can be expressed as a mixture of multiple topics and that the occurrences of words is a result of the topic mixture. Thus, the latent layer of topics that is introduced by pLSA between the image and the tag or visual words, acts as the feature space where both types of words can be com-

bined meaningfully. However, even if these latent topics can be considered to satisfy the requirement of combining the words extracted from heterogeneous modalities without introducing any bias or rendering them meaningless, they still do not solve the problem that, being different expressions of the same abstract meaning, there is a certain amount of dependance between the tag- and visual-words that appear together very frequently. This additional requirement motivates the employment of methods that will allow the cross-words dependencies to influence the nature of the extracted latent topics. In this respect we examine the use of a second level pLSA that treats the latent topics as the observed words. In this way we learn a new pLSA model that allows images to be represented as a vector of meta-topics.

In the following we describe the techniques used to represent an image based on a vocabulary of visual- and tag-words respectively, we provide details on the application of the pLSA model on the extracted word-based image representations, and finally we present an experiment that verifies the efficiency of the presented model in combining heterogeneous types of information.

4.2 Representing images based on the co-occurrence of words

The employment of an efficient image representation scheme that will manage to extract all or most of the important information contained in an image, is a crucial pre-requisite for performing image indexing and retrieval. The scheme adopted in this chapter is based on defining a set of representative "words", that are able to span a sufficiently large portion of the information space that they are used to describe. Then, based on these words each image can be represented with respect to the existence of these words in its content. In the following we describe two techniques for applying the aforementioned general methodology to the spaces of visual content and tags.

4.2.1 Visual-word co-occurrence image representation

In order to represent the visual information carried by an image using a set of visual words, we need to build a bag-of-words representation for the visual content of the images. For the purposes of our work we have used the visual representation scheme adopted in [10] that consists of the following 3 steps: a) the Difference of Gaussian filter is applied on the gray scale version of an image to detect a set of key-points and scales respectively, b) the Scale Invariant Feature Transformation (SIFT) [29] is computed over the local region defined by the key-point and scale, and c) a Visual Word Vocabulary (Codebook) [37] is created by applying the k-means algorithm to cluster in 500 clusters, the total amount of SIFT descriptors that have been extracted from all images. Then, using this Visual Word Vocabulary we are able to vector quantize the SIFT descriptor of an interest point against the set of visual words. This is done by mapping the SIFT descriptor to its closest visual word (i.e., the cluster with minimum distance from its center among all 500 generated clusters) and increasing the corresponding word count. By doing this for all key-points found in an image, the resulting 500-dimensional representation is the visual-word co-

occurrence vector of this image and holds the counts of the occurrences of visual words in its content.

4.2.2 Tag-word co-occurrence image representation

A similar approach has been adopted for representing the tag information that accompanies an image using a set of tag words. As in the previous case, we need to build a bag-of-words representation of the textual content of the image. However, since tags have clear semantics, in this case there is no need to employ clustering for determining which words should be included in the Tag Word Vocabulary (Codebook). Instead, from a large volume of utilized tags we need to select the ones with minimum level of noise and maximum usage by the users. For the purposes of our work we have used the Tag Word Vocabulary constructed by [10] using the following steps. 269,648 images were downloaded from flickr along with their accompanying tags. Among the total set of 425,059 unique tags that have been used by the users to tag these images, there are 9,325 tags that appear more than 100 times. Many of these unique tags arise from spelling errors, while some of them are names etc, which are meaningless for general image annotation. Thus, all these 9,325 unique tags were checked against the WordNet Lexical Database [15] and after removing those tags that do not exist in WordNet, a list with 5,018 unique tags was determined. For the purposes of our work, out of the 5,018 unique tags the first 1,000 that were used most frequently were selected to form the Tag Word Vocabulary. Using this Tag Word Vocabulary we obtain for each image a 1000-dimensional tag-word co-occurrence vector that holds the counts of the occurrences of tag words in its associated set of tags.

4.3 Representing images using aspect models

As already mentioned the goal of pLSA is to introduce a latent (i.e., unobservable) topic layer between images and words. Let us denote $D = \{d_1, \ldots, d_N\}$ the set of images and $W = \{w_1, \ldots, w_M\}$ the set of words. The key idea is to map high-dimensional word count vectors, as the ones described in Section 4.2, to a lower dimensional representation in a so-called latent semantic space [21]. pLSA is based on a statistical model which has been called aspect model [20]. The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in Z = \{z_1, \ldots, z_K\}$ with each observation as shown in Fig. 5(a). A joint probability model over the set of images D and the set of words W is defined by the mixture:

$$P(d,w) = P(d)P(w|d), \quad P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{1}$$

where $P(d)$ denotes the probability of an image to be picked, $P(z|d)$ the probability of a topic given a current document, and $P(w|z)$ the probability of a word given a topic.

Once a topic mixture $P(z|d)$ is derived for an image d, we have a high-level representation of this image with less dimensions from the initial representation that

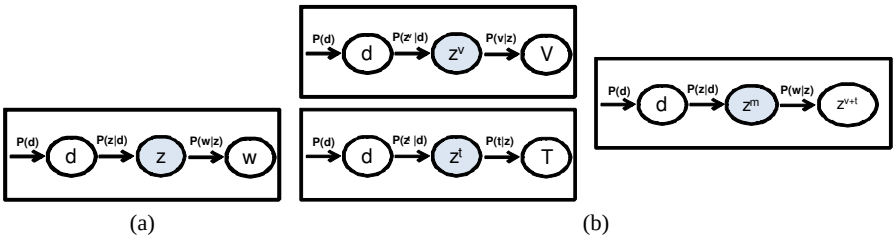


Fig. 5 Image representation using: a) the standard pLSA model, b)the second level pLSA model.

was based on the co-occurrence of words. This is because we commonly choose the number of topics K to be much smaller than the number of words so as to act as bottleneck variables in predicting words. The resulting $K-$dimensional topic vectors can be used directly in a query-by-example image retrieval scenario, if we measure the similarity between two images by computing the distance (e.g., L_1, Euclidean, cosine) between their topic vectors.

For the purposes of our work the pLSA model was independently applied in both visual and tag information space in order to express the images as a mixture of visual and tag topics, respectively. More specifically, in the visual information space the visual-word co-occurrence vectors of all training images were used to train the pLSA model. Then, this model was applied to all testing images in the dataset to derive a new vector representation for each image. In this new representation the vector elements denote the degree to which an image can be expressed using a certain visual topic. Similarly in the tag information space, the tag-word co-occurrence vectors of all training images were used to train a pLSA model, that was used to derive a new vector representation for all testing images. In this case the vector elements denote the degree to which an image can be expressed using a certain tag topic. 100 topics has been used in both cases resulting in two 100-dimensional vectors for each image.

Motivated by the fact that both topic vectors refer to the so-called latent semantic space and express probabilities (i.e., the degree to which a certain topic exists in the image), we assume that the topics obtained from both modalities are homogeneous and can be indiscriminately considered as the words of a common Topic Word Vocabulary. Based on this assumption an image representation that combines information from both modalities can be constructed by concatenating the two 100-dimensional topic vectors. Alternatively, and in order to exploit the cross-words relations between tags $T = \{t_1, \ldots, T_L\}$ and visual words $V = \{v_1, \ldots, v_Q\}$, we can treat the visual- and tag-topics as the observed words for learning a second level pLSA model. This model allows an image to be represented as a vector of meta-topics as depicted in Fig. 5(b). For the purposes of our experiment we have selected the number of meta-topics to be 100, and for every testing image we obtained an 100-dimensional vector. The elements of this vector denote the degree to which an image can be expressed using a certain meta-topic.

4.4 Experimental study

In the following we evaluate the efficiency of different feature spaces for performing image indexing and retrieval. Initially we create an index for testing all images using the corresponding feature space. Then, we use a small subset of the indexed images to perform queries on the index. Performance assessment is based on the relevance between the query and the retrieved images. In our experiment we compare the performance between 5 different feature spaces which are formulated by: a) tag-words, b) visual-words, c) the straightforward concatenation of both types of words, d) topics obtained by concatenating the output of the pLSA model applied on visual- and tag-words, and e) meta-topics obtained by applying a second-level pLSA model on the previously extracted topics.

4.4.1 Test-bed

Our experiment has been conducted on the NUS-WIDE dataset[3] that was created by the NUS's Lab for Media Search [10]. The dataset contains 269,648 images that have been downloaded from flickr together with their tags. For all images the authors released 500-dimensional co-occurrence vectors for visual words (as described in Section 4.2.1), as well as 1000-dimensional co-occurrence vectors for tag-words (as described in Section 4.2.2). Moreover, the ground-truth for 81 concepts has been provided to facilitate evaluation. For the purposes of our evaluation we have used a subsample of 20,000 images for training (I^{train}) and 20,000 for testing (I^{test}). I^{train} has been used for training the pLSA models and I^{test} to create the index and perform image retrieval. Finally, 1,000 images from I^{test} were selected to act as queries I^{query}.

4.4.2 Evaluation metric

Since we are considering an image retrieval scenario, the Average Precision (AP) metric was selected to assess the performance of the different feature spaces. AP favors the algorithms that are able not only to retrieve the correct images, but to retrieve them as earlier as possible in a ranked list of retrieved results. Thus, for two algorithms that retrieve the same amount of correct images, although achieving similar performance with respect to recall and precision, the one that retrieves the correct images earlier in the ranked list of results will exhibit highest score for AP. Average precision is expressed by the following equation.

$$AP = \frac{\sum_{r=1}^{N} Pr(r) \cdot rel(r)}{\#\text{ relevant images}} \tag{2}$$

where r is the current rank, N is the number of retrieved images, $rel()$ is a binary function that determines the relevance of the image at the given rank with the query image. $rel()$ outputs 1 if the image in the given rank is annotated with at least one concept in common with the query image and 0 otherwise. $Pr(r)$ is the precision at the given rank and is calculated by:

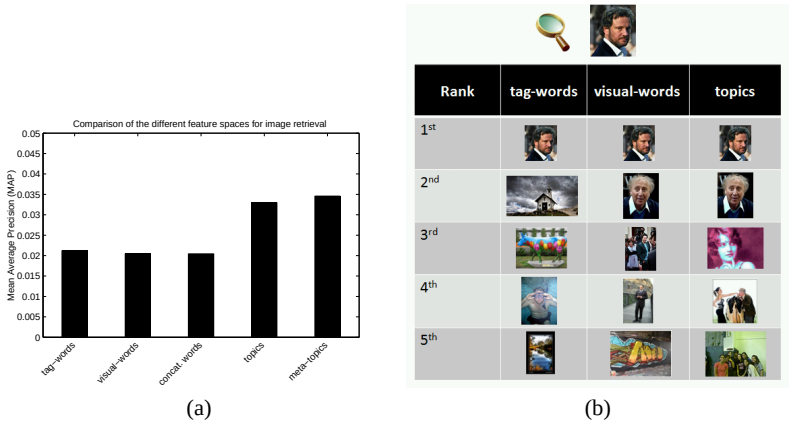
[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

$$Pr(r) = \frac{\text{\# relevant retrieved images of rank r or less}}{r} \tag{3}$$

AP measures the retrieval performance of the method using one image as query. In order to obtain one global performance score for all 1,000 images that we were used as queries, we employed the Mean Average Precision (MAP). MAP is the mean of AP scores over a set of queries. In our experiment the MAP was calculated over the 1,0000 query images included in I^{query}.

4.4.3 Results

Fig. 6(a) depicts the MAP scores for all evaluated feature spaces. We notice that tag-words and visual-words exhibit almost similar performance, showing that both modalities carry equally important information for retrieving relevant images. As expected, the straightforward combination of both modalities by simply concatenating their word count vectors, fails to combine them efficiently and performs slightly worse than the uni-modal cases. This is not the case for the space of latent topics obtained using pLSA. We note a significant increase of the retrieval performance that verifies the ability of the pLSA-generated space to efficiently combine information from heterogeneous sources. Finally, the space of meta-topics manages to further improve the retrieval performance by a small amount, showing that the exploitation of cross-word relations can only benefit the resulting feature space. Fig. 6(b) shows indicative retrieval retrieval examples using tag-words, visual-words and topics.



(a) (b)

Fig. 6 a) Comparison diagram between 5 different feature spaces for image indexing and retrieval, b) Indicative retrieval examples using tag-words, visual-words, and latent topics

In conclusion, we should stress the great potential of exploiting the information residing across different modalities, provided that we will manage to overcome the problems arising from the heterogeneous nature of sources. The use of aspect mod-

els has been proven to be an efficient method for combining the visual and tag information carried by the images on the Web. However, a similar methodology can be used to incorporate additional modalities such as geotagging or user-related information. In the end, we should also highlight the existence of dependencies between the information carried by different modalities, and the need to investigate methods that will manage to exploit these dependencies. By doing so, additional means can be discovered for boosting the combination efficiency of multi-modal data.

5 Conclusions

In this chapter we have discussed the importance of multi-modal analysis in social media. Given the widespread adoption of social networks and web 2.0, there is a major trend in the new digital landscape towards the interconnection of platforms, networks and most importantly data. Different types of information are associated with the same digital resource expressing different aspects of its meaning (e.g., textual and visual descriptions, spatio-temporal information, popularity, etc). The existence of techniques that will manage to exploit all these aspects is crucial for successfully adopting to this new digital landscape. Initial works on multi-modal analysis have shown that the straightforward application of well-studied uni-modal methods are rarely beneficial in multi-modal settings. Problems arise from the heterogeneity of sources and the very different characteristics of the data. In this chapter we have presented two methods that alleviate the aforementioned problems in two scenarios involving tagged images.

The ant-inspired algorithm presented in Section 3 formulates multi-modal analysis as a feature selection problem. Feature weights are used to favor the modality that best describes the analyzed image and image clustering is performed using the corresponding vectors of weighted features. Thus, depending on the calculated weights, the resulting clusters may refer to different feature subspaces, emphasizing on different aspects of the image content. The method presented in Section 4 emphasizes on the need to discover techniques that will manage to exploit the relations existing across modalities, in cases where two or more of the modalities constitute different expressions of the same abstract meaning. Using the presented second level pLSA model the extracted hierarchy of topics manages to exploit the cross-modal relations between the visual and tag features and improve the retrieval performance. Both methods demonstrate their superiority against their uni-modal counterparts and underline the importance of efficient and effective methodologies for the optimal combination of multi-modal data.

As avenues for potential research we can refer to the need of investigating methods that will manage to efficiently combined a considerably higher number of modalities. Although many of the presented techniques claim that can be naturally scaled to support the combination of more than two modalities, only few of them have been evaluated with real world data. The need for such methods is particularly motivated by the constantly increasing level of information flow between different social networks, which leverages the multi-source aspect of social media. For instance an image in flickr[4] can be bookmarked in delicious[5] or tagged using an event from last.fm[6], allowing its association with information from multiple sources. This information flow can be also the motive for discovering new application scenarios that will be able to exploit the added value offered by performing multi-modal analysis on the exchanged data.

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

[5] http://www.delicious.com/

[6] http://www.last.fm

Acknowledgements This work was sponsored by the European Commission as part of the Information Society Technologies (IST) programme under grant agreement n215453 - WeKnowIt and the contract FP7-248984 GLOCAL.

References

1. MPEG-7 Visual Experimentation Model (XM). Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4062 (2001)
2. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the ACM SIGMOD Int'l Conference on Management of Data, Seattle, Washington, pp. 94–105. ACM Press (1998)
3. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data. Data Mining and Knowledge Discovery **11**, 5–33 (2005)
4. Aurnhammer, M., Hanappe, P., Steels, L.: Augmenting navigation for collaborative tagging with emergent semantics. In: International Semantic Web Conference (2006)
5. Becker, H., Naaman, M., Gravano, L.: Event identification in social media. In: 12th International Workshop on the Web and Databases, WebDB (2009)
6. Becker, H., Naaman, M., Gravano, L.: Learning similarity metrics for event identification in social media. In: WSDM '10: Proceedings of the third ACM international conference on Web search and data mining, pp. 291–300. ACM, New York, NY, USA (2010)
7. Blum, C.: Ant colony optimization: Introduction and recent trends. Physics of Life Reviews **2**, 353–373 (2005)
8. Caro, G.D., Ducatelle, F., Gambardella, L.M.: Anthocnet: an adaptive nature-inspired algorithm for routing in mobile ad hoc networks. European Transactions on Telecommunications **16**(5), 443–455 (2005)
9. Cheng, C.H., Fu, A.W., Zhang, Y.: Entropy-based subspace clustering for mining numerical data. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '99, pp. 84–93. ACM, New York, NY, USA (1999)
10. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval, pp. 1–9. ACM, New York, NY, USA (2009). DOI http://doi.acm.org/10.1145/1646396.1646452
11. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th international conference on World wide web, WWW '09, pp. 761–770. ACM, New York, NY, USA (2009)
12. Domeniconi, C., Al-Razgan, M.: Weighted cluster ensembles: Methods and analysis. ACM Transactions on Knowledge Discovery from Data **2**, 17:1–17:40 (2009)
13. Dorigo, M.: Optimization, Learning and Natural Algorithms. Ph.D. thesis, Politecnico di Milano, Italy (1992)
14. Dorigo, M., Caro, G.D.: The ant colony optimization meta-heuristic (1999)
15. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
16. Franz, T., Schultz, A., Sizov, S., Staab, S.: Triplerank: Ranking semantic web data by tensor decomposition. In: ISWC '09: Proceedings of the 8th International Semantic Web Conference, pp. 213–228. Springer-Verlag, Berlin, Heidelberg (2009)
17. Giannakidou, E., Kompatsiaris, I., Vakali, A.: Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems. In: ICSC, pp. 128–135 (2008)
18. Giannakidou, E., Koutsonikola, V.A., Vakali, A., Kompatsiaris, Y.: Co-clustering tags and social data sources. In: WAIM, pp. 317–324 (2008)
19. Harshman, R.A., Lundy, M.E.: Parafac: Parallel factor analysis. Computational Statistics & Data Analysis **18**(1), 39 – 72 (1994)
20. Hofmann, T.: Unsupervised learning from dyadic data. pp. 466–472. MIT Press (1998)
21. Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, UAI'99. Stockholm (1999). URL citeseer.ist.psu.edu/hofmann99probabilistic.html
22. Kennedy, L., Naaman, M.: Less talk, more rock: automated organization of community-contributed collections of concert videos. In: Proceedings of the 18th international conference on World wide web, WWW '09, pp. 311–320. ACM, New York, NY, USA (2009)

23. Kennedy, L.S., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections. In: ACM Multimedia, pp. 631–640 (2007)
24. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review **51**(3), 455–500 (2009). DOI 10.1137/07070111X
25. Lathauwer, L.D., Moor, B.D., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)
26. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: MULTIMEDIA '03, pp. 604–611. ACM, New York, USA (2003)
27. Lienhart, R., Romberg, S., Hörster, E.: Multilayer plsa for multimodal image retrieval. In: CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval, pp. 1–8. ACM, New York, NY, USA (2009). DOI http://doi.acm.org/10.1145/1646396.1646408
28. Lindstaedt, S., Pammer, V., Mörzinger, R., Kern, R., Mülner, H., Wagner, C.: Recommending tags for pictures based on text, visual content and user context. In: Proceedings of the 2008 Third International Conference on Internet and Web Applications and Services, pp. 506–511. IEEE Computer Society, Washington, DC, USA (2008)
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
30. Magalhaes, J., Rüger, S.: Information-theoretic semantic multimedia indexing. In: CIVR '07, pp. 619–626. ACM, New York, USA (2007). DOI http://doi.acm.org/10.1145/1282280.1282368
31. Manjunath, B.S., Ohm, J.R., Vinod, V.V., Yamada, A.: Colour and texture descriptors. IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7 **11**(6), 703–715 (2001)
32. Olivares, X., Ciaramita, M., van Zwol, R.: Boosting image retrieval through aggregating search results based on visual annotations. In: Proceeding of the 16th ACM international conference on Multimedia, MM '08, pp. 189–198. ACM, New York, NY, USA (2008)
33. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. SIGKDD Explor. Newsl. **6**, 90–105 (2004)
34. Piatrik, T., Izquierdo, E.: Subspace clustering of images using ant colony optimisation. In: 16th IEEE International Conference on Image Processing (ICIP), pp. 229 –232 (2009)
35. Quack, T., Leibe, B., Gool, L.J.V.: World-scale mining of objects and events from community photo collections. In: CIVR, pp. 47–56 (2008)
36. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceeding of the 17th international conference on World Wide Web, WWW '08, pp. 327–336. ACM, New York, NY, USA (2008)
37. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, p. 1470. IEEE Computer Society, Washington, DC, USA (2003)
38. Sizov, S.: Geofolk: latent spatial semantics in web 2.0 social media. In: WSDM '10: Proceedings of the third ACM international conference on Web search and data mining, pp. 281–290. ACM, New York, NY, USA (2010). DOI http://doi.acm.org/10.1145/1718487.1718522
39. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Tag recommendations based on tensor dimensionality reduction. In: RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems, pp. 43–50. ACM, New York, NY, USA (2008)
40. Wu, Y., Chang, E.Y., Chang, K.C.C., Smith, J.R.: Optimal multimodal fusion for multimedia data analysis. In: MULTIMEDIA '04, pp. 572–579. ACM, New York, USA (2004)
41. Xu, R., Wunsch, I.: Survey of clustering algorithms. Neural Networks, IEEE Transactions on **16**(3), 645–678 (2005)