Deriving Taxonomies from Automatic Analysis of Group Membership Structure in Large Social Networks.

Marc Egger, Kai Fischbach, Peter Gloor, Andre Lang, Mark Sprenger

Seminar für Wirtschaftsinformatik und Informationsmanagement
Universität Köln
Pohligstraße 1
50969 Köln

egger@wim.uni-koeln.de
fischbach@wim.uni-koeln.de
pgloor@mit.edu
lang@wim.uni-koeln.de
sprenger@wim.uni-koeln.de

Abstract: We develop a method to create taxonomies in large social networks solely based on users group membership information. We illustrate our technique using an example of the Flickr photo sharing network. This photo sharing community is a social network that enables users to collectively store digital pictures, interact and form groups of interest. Based on our earlier research of success indicators for individual actors on the Flickr Community, we extend the focus to groups of interest. Introducing the metric "GroupConnectivity" we perform a community segmentation. We then develop a method to automatically create taxonomies without the need for folksonomies or pre-existing ontologies.

1 Introduction

Currently there is a rapidly growing number of "social networks", "online communities" and other platforms that enable Internet users to participate in the evolution of the particular Website. Most were established under the umbrella of "Web 2.0" which has become a synonym for Internet applications that enable user participation. These online "social networks" include three parts as standard properties. First, its users have lists of other users they exchange information with. Second, a single site deals with one particular topic. Examples can be information exchange with neighbours, colleagues and friends or sharing of content like video, audio or photos. Third, users can form groups of interest where any user is able to create a new group other Website members can join. Based on earlier research in [EFGLS08] where we developed metrics to identify successful actors using information from the first two categories introduced above – contact lists and single focus, we now focus on the third – group membership information. In the first step this information helps to visually segment the Web site's content. In the second step, using the resulting linkage between groups, we develop a method to create taxono-

mies of related groups solely from the degree of shared membership and group size. Both methods are applied to the data set of the Flickr community that was gathered between November 2007 and January 2008, covering about 35% of the community. The techniques introduced in this paper are easily applicable to other networking sites.

Previous research in this area was made. Since the introduction of the public Flickr API several groups have started exploring the Flickr network. Prieur et al. [Pr08] gathered the entire Flickr user base in 2006 and made a thorough analysis on the 50.000 most active users by means of photos, contacts, favourites and comments. They distinguished three different ways of using Flickr. Furthermore, they analysed 450 random Flickr groups with similar size and compared their social density versus the similarity of tags on photos posted by the group members, showing a high diversity in both. Negoescu and Gatica-Perez [NG08a] presented a statistical analysis of 51,000 groups, showing the behavioural patterns of users when sharing photos with groups. Based on aggregated group tags, they developed a topic-based representation model of groups. This model was proposed as a basis for searching expert groups on certain topics and also for visually browsing through thematically adjacent groups. They extended this idea to users' tags [NG08b], being able to search and navigate through groups and users alike.

Some effort has been made to use tags in order to build up taxonomies. Schmitz [Sc06] made a first attempt in using a co-occurrence model on a Flickr tag set to find subsumptions in order to build up a taxonomic ontology. Others tried to match tags to predefined ontologies like WordNet [LEC07]. Van Damme, Hepp and Siorpaes [DHS07] gave an overview of the obstacles when processing folksonomies. They suggested a mixed approach of statistical analysis and mapping to lexical resources or existing ontologies. Nevertheless, to our knowledge no one has used the network connectivity itself so far, i.e. the users and the groups they are in, as a source for building up taxonomies.

The paper is divided into six sections. In this section we focussed on related literature and the research question. In section 2 we describe the metric used for the analysis. Section 3 deals with the application of the metric on the underlying data set for community segmentation. Section 4 builds on the segmentation and metric to develop a method for creating taxonomies without the use of content-related information. Section 5 includes a discussion on the results, while section 6 summarizes the work and gives suggestions for future research.

2 Metric for co-membership of users in different groups.

During earlier research on group memberships of individual users in [EFGLS08] we noticed that users often join different groups of related topics. For example, individuals who are member of groups that deal with "black and white photography" often also join groups on "sepia photography". To test if this assumption holds true, we introduce "GroupConnectivity", a metric for the co-membership of users in different groups. We will demonstrate its usefulness in section 3 and 4.

Based on the fraction of users who are in two groups of interest at the same time, it describes how close these two groups are connected. Mathematically it can be described as the intersection of two sets of users.

$$\text{GroupConnectivity} = \frac{\text{number of users that are in both groups}}{\min\ (\text{number of members(group A, group B)})}$$

Equation 2-1

The metric is defined in an interval between 0 and 1, whereas 1 means that every user who is member of the smaller group is also member of the larger group. Figure 2-1 clarifies the concept using an example of two groups having 5 and 3 members.

This metric has the advantage that is not influenced by the size of each of the groups or scale between them, as it is always defining the maximum possible connectivity between any kind of groups as 1.

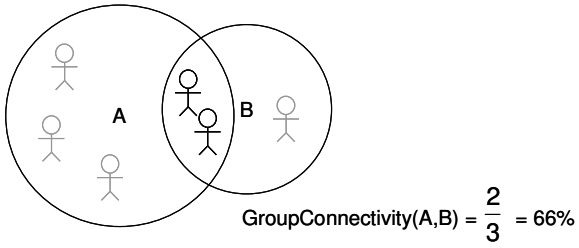

Figure 2-1: Example of GroupConnectivity metric

3 Segmenting the community by means of group memberships

To segment the community, we now use the GroupConnectivity metric and the tool Condor[1] to visualize the resulting structure. The visualization is based on a force model where the layout of nodes and edges is determined by a spring embedder. Therefore nodes repel each other while edges act as springs. As stated in section 2, GroupConnectivity represents the number of users that are member of two groups at the same time. Because users who are interested in a certain topic usually join several groups related to it, GroupConnectivity can be considered as a degree of thematic similarity between these groups.

Our dataset consists of 3.6 million users. 34% of these are members of at least one group, making up a total number of 18.3 million group memberships within 160.000 groups. For the analysis we choose a set of the 300 best-performing groups in Flickr. The performance of groups was measured by the GroupFavCount metric that we define as the sum of FavCount of all group members. The FavCount introduced in [EFGLS08] is based on how often other users have added pictures of a particular user to their list of favourites. A first view on our data revealed that none of the potential combinations of groups has a GroupConnectivity of zero. For that reason, adding an edge for each

[1] Developed by galaxyadvisors AG, http://www.galaxyadvisors.com

GroupConnectivity larger than zero would result in an unclear, complete graph. Therefore we sorted out connections of lower GroupConnectivity value and visualise only connections exceeding a certain threshold. Figure 3-1 shows the GroupConnectivity's frequency distribution, which has its peak at 11%. We apply a threshold of 50% while using the value of GroupConnectivity as the weight of the edges. The threshold decreases the total number of visible edges. This is needed in order to get a meaningful visualization, as drawing all edges with a GroupConnectivity >0 would result in an almost complete graph. Hiding the less important edges gives a better insight into the structure.

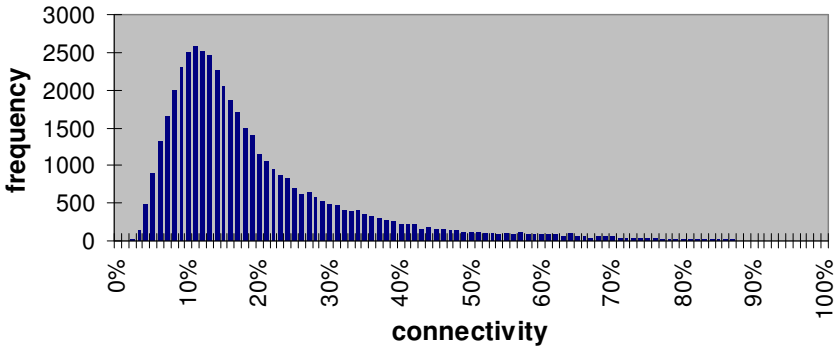

Figure 3-1: Histogram of GroupConnectivity

The resulting graph reveals the structure of Flickr's group networks (Figure 3-2). FlickrCentral, the biggest group, is located in the middle of the network, serving as a central hub between the different parts of the network. Taking a closer look at Figure 3-3 and Figure 3-4 we verified that the assumption of groups being related to each other by overlapping sets of members having similar topics was reasonable.

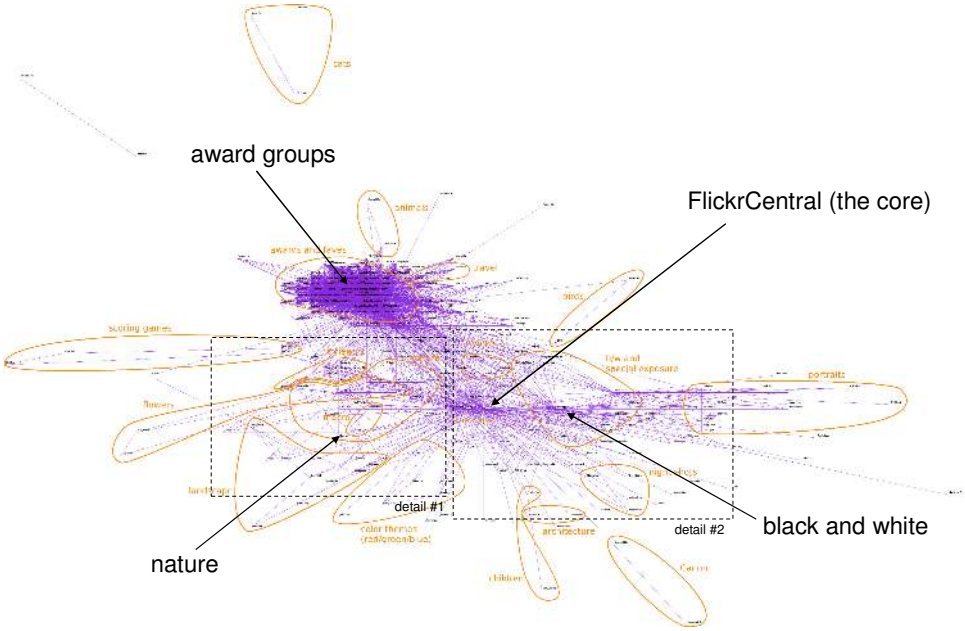

Figure 3-2: Flickr GroupConnectivity graph overview

The group “nature” is located in the centre of detail #1 (Figure 3-3). Assuming that the group’s name corresponds to its topic, the clusters close to this group are also thematically related. Around the group “nature” we find groups that bear names like “Sunsets”, “Sunrise”, “Beautiful”, “FLOWERS”, “FLOWERPIX”, “Flower pow”, “Flower Mac”, “Trees”, “All Animal”, “Sky”, “NATUREPIX”.

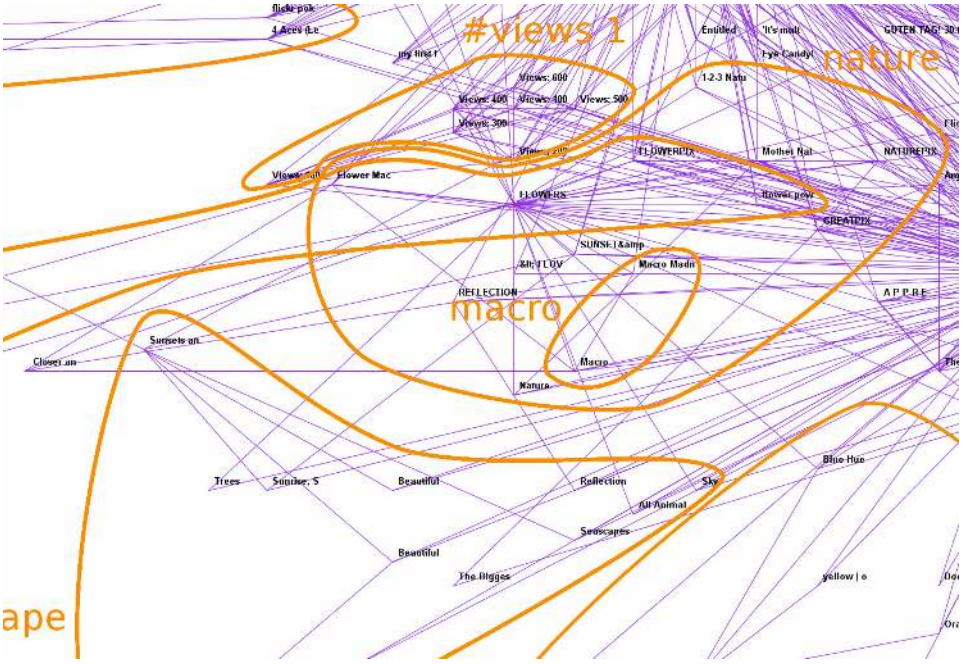

Figure 3-3: Flickr GroupConnectivity graph, detail #1

Zooming into detail #2 (Figure 3-4) reveals the same results. Groups with names that suggest related topics arrange themselves in clusters around the original topic respectively group. Here names like “Sepia”, “Lights and Shadow”, “Silhouette” appear around the group “Black and White” and groups like “Picturing”, “Portraits”, “all People”, “Faces” and “Portrait” form a cluster, just to name a few examples.

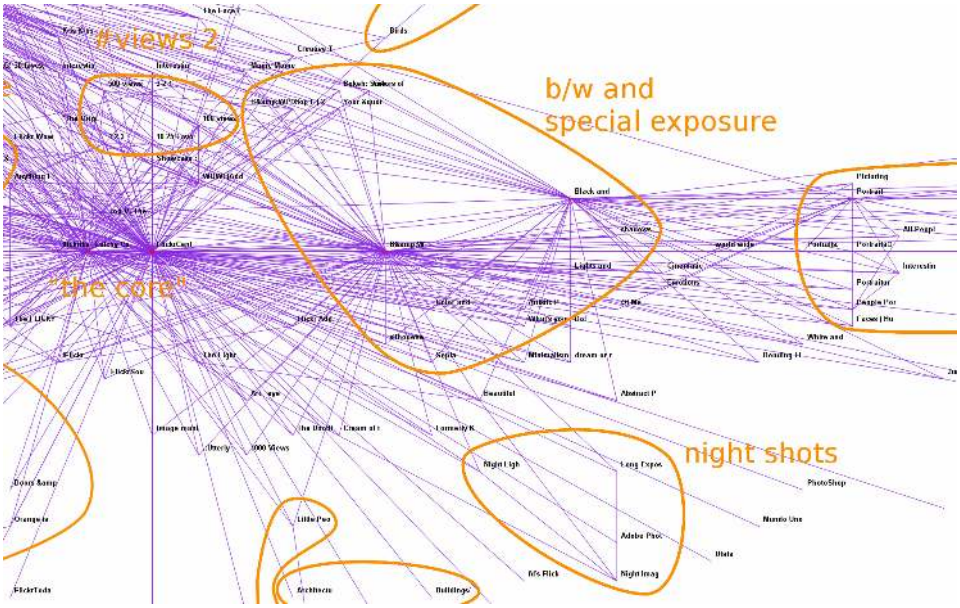

Figure 3-4: Flickr GroupConnectivity graph detail #2

4 Creating taxonomies based on GroupConnectivity

Following the community segmentation, we now develop a method to organise social network groups into a hierarchic tree. The idea is based on two observations made during the development of the previous results. First, GroupConnectivity is high between groups of similar interest. We were able to show this in the visualisation by having groups of similar interest located in immediate vicinity to each other. Second, it seems that larger groups usually cover broader, more general topics than smaller groups[2]. We will now provide an algorithm, which builds a semantic hierarchy, i.e. a taxonomy of the groups by using the GroupConnectivity metric.

By definition, the largest group in the network becomes the root node. After that, we find a parent for each remaining group A. Among all other groups that are larger than A by member size, the parent of A will be the one that has the highest GroupConnectivity in relation to currently examined group A.

[2] E.g. the "Catchy Colors" group is larger than the "blue hue" group.

The pseudocode is shown in Figure 4-1. Figure 4-2 gives an example for the selection of the parent node. The result is a tree with descending group sizes, descending from the root to the leaves.

```
for each group A do
    – within all groups that are larger than A
    – find a group B with GroupConnectivity(A,B)→max!
    – → A is child of B
```

Figure 4-1: Hierarchic tree algorithm (pseudocode)

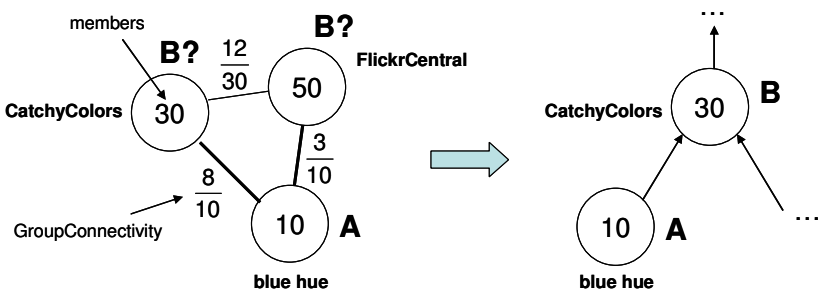

Figure 4-2: Example of the hierarchic tree algorithm

We applied the algorithm to the same dataset of 300 top performing groups which we already used before and visualized the resulting tree. The tree depicted in Figure 4-3 shows the expected structure. The root node, “Flickr central” is at the far right. The larger and more general groups are parents of the smaller and more specific groups. Moreover, the parent groups represent more generic terms while dealing with the same topic. For example, the "flowers" group is parent of many smaller groups dealing with different kinds of flower pictures, such as "flower macro", "flowerpix", while "Catchy Colors" is parent of several groups devoted to a single colour. Therefore, the algorithm has succeeded in automatically creating a taxonomy of the group topics without using further content-related information. In a few cases, our algorithm failed in creating an intuitive general-detail relationship. This happens when no larger group covering the more general topic exists. For example, as no “plant” group exists, the “tree” group becomes a child of the larger “flower” group, instead of both being children of a “plant” group. In other cases, such as the “what’s your angle, buddy?” and “catchy colors” groups, relations show up between obviously unrelated topics as represented by group names, although the content may still be closely related.

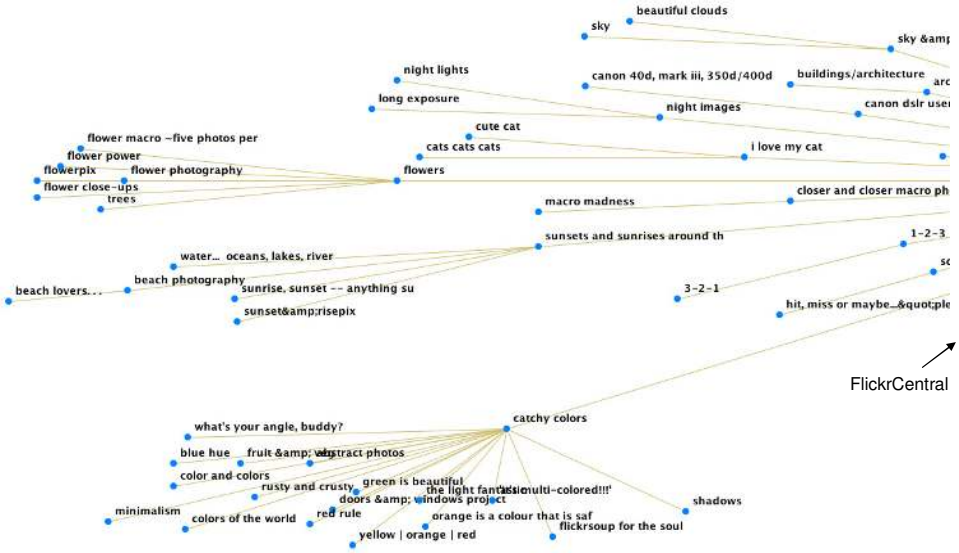

Figure 4-3: Hierarchic tree of analysed group set, detail

Due to the simplicity and universality of this algorithm, it is applicable to other social networks as well. The only input data needed is the group memberships of the social network's users. Based on this information, one can derive the number of group members as well as the GroupConnectivity in a first pre-processing step.

It is remarkable that no tags or similar information provided by users or editors is needed. The structure is built "by itself", without any semantic input, but creating a hierarchy that holds a semantic meaning.

5 Discussion

Structuring, classification and segmentation of large social networks is a difficult, yet often indispensable task. To discover interesting clusters, creating a hierarchic sitemap or finding thematically adjacent topics requires efficient ways to build up a classification system. Folksonomies and editorial systems are ways to perform this task, although they require at least a certain amount of human intervention for assigning the tags or classifying the content. This may be perfectly acceptable for either networks that rarely change, making the classification a one-time task; or where human resources needed to assign tags are freely available. Nevertheless, provided that no or insufficient tag information is available, this task is prohibitively expensive for large networks which grow or change rapidly. This is a common characteristic of social networks.

Usual taxonomy creation approaches are based on user generated classification information respectively folksonomies and statistical analysis or matching to predefined ontolo-

gies. The resulting taxonomy either consists of concepts defined by a number of more or less adequate tags [Sc06] or a concept in a predefined hierarchy [LEC07]. Dealing with folksonomies lead to problems of content processing such as synonyms, homonyms and other linguistic related challenges. When using folksonomies as a basis for a taxonomy, the sum of tags might not be sufficient to catch the concept that the folksonomies aim to describe.

Our alternative approach is applicable when no content information is available. Since the group name serves as the concept description, no content processing issues arise. In some cases these group names may even be more descriptive and precise than tags or given categories can be.

In the following we try to explain why the GroupConnectity leads to intuitively reasonable results. Arguing from the evolutionary process of a social network, one could conclude that users who are members of a more general group tend to create a new group if the topic they are interested in is too specific to be covered by the general group. Another explanation could be that users join groups of similar topics, which gives groups with similar topics a higher GroupConnectivity. Given that larger groups are covering more general topics, it is only natural that the larger group with highest GroupConnectivity is the parent node inside a hierarchic tree and covers a more general topic.

From a practical perspective, the hierarchic tree algorithm is a cheap and uncomplicated way to get a first tree diagram of the network's hierarchy. It can be used to build up a sitemap of a social network, giving editors a first impression of what the categories should look like and providing them with a basis to start from. A further application is using it as a preliminary filter for the social network analysis of large networks. Visual representations of large social networks can suffer when too many nodes are displayed at the same time. This is especially true for densely connected networks, where detail information may get lost. The algorithm introduced in this paper is able to create an intuitive overview of the network from which the relevant nodes and subtrees containing topics of close vicinity can easily be selected beforehand.

Because of its low resource requirements, it is also conceivable to use the algorithm on many other, completely different kinds of networks. For example, when having the trails of users on different Websites, these Websites can be seen as "groups" as well, counting each visit of a user as a "group membership". Processing this information with our algorithm could result in a hierarchic tree classifying all considered Websites, similar to hierarchic directories like the Yahoo! Directory[3] or the Open Directory Project[4].

6 Summary and Outlook

Based on co-membership of users in different groups we were able to develop a metric that helps to determine clusters of related topics represented by group names. These findings were used for automated clustering of topics that are thematically near to the

[3] Yahoo Inc., Yahoo! Directory, http://dir.yahoo.com/
[4] Netscape Communications Corporation, Open Directory Project, http://www.dmoz.org/

groups a user is already a member of. This prerequisite is also used to build a taxonomy tree that hierarchically organizes the groups from general to specific related topics.

Further research is needed to evaluate and validate our approach against existing taxonomies, for example by comparing to a taxonomy resulting from manual card sorting [URK01].

Beyond that, improvements and variations of the algorithm are possible. We successfully implemented "penalty" rules to the GroupConnectivity selection criteria. For example, it is possible to prevent that many very small nodes get attached to the root node by giving a penalty to combinations of very small and very large groups. This forces the small nodes to get attached deeper in the tree. By using these penalties, it is also possible to influence the depth and width of the resulting trees. We are currently working to develop new applications and further optimizations for our algorithm.

Nevertheless, this paper shows that a community literally has the capability to shape and maps its own content.

7 Bibliography

[DHS07] Van Damme, C.; Hepp, M.; Siorpaes, K: An Integrated Approach for Turning Folksonomies into Ontologies, Proceedings of the ESWC 2007 Workshop "Bridging the Gap between Semantic Web and Web 2.0", Innsbruck, Austria, 2007. CEUR, 2007; pp. 57-70.

[EFGLS08] Egger, M.; Fischbach, K.; Gloor, P.; Lang, A.; Sprenger, M.: How to Identify Successful Actors of the Flickr Community and How to Determine Their Attributes, Informatik 2008 - Beherrschbare Systeme - Dank Informatik. Series: Lecture Notes in Informatics, no. 134, pp. 947-952.

[LEC07] Laniado, D.; Eynard, D.; Colombetti, M.: Using WordNet to turn a folksonomy into a hierarchy of concepts. Fourth Italian Workshop on Semantic Web Applications and Perspectives (SWAP 2007), Bari, Italy, 2007; pp. 192-201. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.4393&rep=rep1&type=pdf#page=200

[NG08a] Negoescu, R.-A.; Gatica-Perez, D.: Analyzing Flickr Groups. Proceedings of the 2008 international conference on Content-based image and video retrieval, Niagara Falls, Canada, 2008. ACM Press, New York, 2008; pp. 417-426.

[NG08b] Negoescu, R.-A.; Gatica-Perez, D.: Topickr: Flickr Groups and Users Reloaded. Proceeding of the 16th ACM international conference on Multimedia, Vancouver, Canada, 2008. ACM Press , New York, 2008; pp. 857-860.

[Pr08] Prieur, C.; Cardon, D.; Beuscart, J.-S.; Pissard, N.; Pons, P.: The Strength of Weak cooperation: A Case Study on Flickr. arXiv, 2008. [Online]. Available from: http://uk.arxiv.org/abs/0802.2317

[Sc06] Schmitz, P.: Inducing Ontology from Flickr Tags. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, 2006. Available from: http://www.ibiblio.org/www_tagging/2006/22.pdf

[URK01] Upchurch, L.; Rugg, G.; Kitchenham, B. 2001. Using card sorts to elicit Web page quality attributes. IEEE Software 18(4), 2001. IEEE Computer Society Press, Los Alamitos, CA, USA, 2001; pp. 84-89.