

# On the Feasibility of a Tag-based Approach for Deciding Which Objects a Picture Shows: An Empirical Study

Viktoria Pammer<sup>1</sup>, Barbara Kump<sup>2</sup>, and Stefanie Lindstaedt<sup>1</sup>

<sup>1</sup> Know-Center {vpammer,slind}@know-center.at

<sup>2</sup> Knowledge Management Institute, TU Graz bkump@tugraz.at \*

**Abstract.** Many online platforms allow users to describe resources with freely chosen keywords, so called tags. The specific meaning of a tag as well as its specific relation to the tagged resource are left open for interpretation to the user. Although human users mostly have a fair chance at interpreting it, machines do not. An algorithmic approach for identifying descriptive tags however could prove useful for intelligent search for pictures and providing first-cut overviews over tagged picture repositories. In this paper we investigate the characteristics of the problem to decide which tags describe visible entities on a given picture. Based on a systematic user study, we are able to discuss in detail the problems involved for both humans and machines when identifying descriptive tags. Furthermore, we investigate the general feasibility of developing a tag-based algorithm tackling this question. Finally, a concrete implementation and its evaluation are presented.

## 1 Introduction

Various social software and collaborative tagging platforms have sprung up everywhere on the web. They enable users to describe photos, news, blogs, research publications and web bookmarks with freely chosen keywords, so called *tags*, and to share both their content and their tags with other users. The appeal of tagging lies in its simplicity for the tag producer, who can attach to a resource any keyword he or she deems appropriate. No explanation of the exact meaning of the tag has to be provided and no rules restricting the vocabulary have to be adhered to. This ease of use on the tag producer's side creates a disadvantage on the side of the tag consumer. The (human or machine) tag consumer who wants to use other people's tags for some purpose has to interpret the given tags. Human tag consumers may often be able to do so, given both the tag and the tagged resource. However, they may already have difficulty in finding the

---

\* The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

appropriate tags to formulate queries [1]. Even more at disadvantage are machines who want to consume tags. Consequently, multiple attempts at enriching the tags themselves with semantics have been made (Sec. 2).

In this paper, we analyse the semantics of a specific relation between tags and pictures, namely the “shows”-relation which expresses that a tag refers to an entity which is visible on the picture. We present a problem analysis (Sec. 3) and discussion (Sec. 5) of the challenge of identifying such a relation, a theoretical upper-limit for tag-based algorithms attempting to identify the “shows”-relation and describe a proof-of-concept implementation and its evaluation (Sec. 4).

## 2 The Semantics of Tags

The abundance of tagged data on the web today has prompted researchers to investigate the semantics of tags. In many cases, the approaches taken are attempts to specify the exact meaning of each tag within one collaborative tagging system. Such approaches typically try deriving ontologies of some kinds from the folksonomic structures relating users, tags and resources which underlie collaborative tagging systems. For instance, Mika [3] derives light-weight ontologies directly from folksonomies, such that they represent a *community’s view* on a topic. A different approach is taken e.g. by Schmitz [4], who uses WordNet to lay an ontology *over* the tags by mapping tags to concepts in WordNet. Rattenbury et al. [5] infer whether a tag describes an event or a location based on characteristics of the temporal and spatial distribution of tags. Considered under the aspect of the semantics of the relation between tags and resources, tags that describe a location or an event do not only fall into different categories, in that they would be put in different places in an ontology, but also inherently differ in the relation they can have to a picture.

More directly related to our research, Golder and Hubermann [1] describe the *kinds of tags* encountered within collaborative tagging systems. According to them, tags can for instance identify who or what the resource is about, the kind of the resource (e.g., article, picture, example), who owns the resource, or characteristics of the resource (e.g., awesome, interesting). The kinds of tags correspond, in a different terminology, to the relation between tags and the resources they are attached to. Note that the starting point for discussing semantics is different when discussing tags in collaborative tagging systems or labels in coordinated image labelling efforts, where the goal is to create training sets for automatic classifiers (see e.g. [2] or the ESP Game<sup>3</sup>). In the latter, the semantics of labels is predefined by the goal, i.e. a correct label describes an object visible on the labelled picture.

Similar to the categorisation of different kinds of tags by Golder and Hubermann [1], Bechhofer [6] categorises different kinds of semantic annotations. semantic annotations are different from tags insofar as they unambiguously define the meaning of each tag (the tag becomes a concept) and the meaning of each relation between a tag and a resource. One of the possible meanings of a relation

---

<sup>3</sup> <http://www.espgame.org/gwap/gamesPreview/espgame/>

cited in the paper by Bechhofer is “instance reference”, to which our contribution is strongly linked.

The core of our research is the observation that not all tags attached to a picture describe the visible content of the picture [1, 6]. In collaborative tagging environments, people tag not only what is visible on a picture (e.g. the Eiffel Tower), but they add additional tags which for instance relate to the context in which the picture was taken (e.g. Paris trip, holiday), to adjectives that describe the picture (e.g., impressive, high), or are just statements to express the user’s likes or dislikes (e.g., wow!). We are specifically interested in the “shows”-relation between a picture and a tag. It can be verbalised as “Picture <X> shows a(n) <tag>” where “<X>” stands for a specific picture, and “<tag>” stands for a specific tag. For instance, one could say “The picture in Figure 1 shows a flower”. For brevity, we often refer to tags in a “shows”-relation to the picture they are related to as “descriptive tags”. The “shows”-relation asserts that there is some part on the picture on which a real-world instance corresponding to the tag is visible, but the exact part is not further specified and does not have a URI. Indeed, it is only the picture as a whole which has a URI. Furthermore, for the time being we do not differentiate between the case where the tag denotes a concept and an instance of this concept is visible on the picture (e.g. “castle”), and the case where the tag denotes an instance and this instance is visible on the picture (e.g. “Versailles”).

As an example, consider Fig. 1. It shows flowers, a garden and a house at the background. There are many more tags assigned to the picture. Depending on the viewer’s knowledge she might recognise the flower as a hollyhock and the house as “Haus Liebermann”<sup>4</sup>. Then, if the viewer spoke German, the tags “Garten” and “Malve” could be recognised as being translations of “garden” and “hollyhock”. It is the ultimate goal of our work to devise an algorithm that can identify precisely this “shows”-relation. Our concrete research question is, **how humans do and machines possibly can decide whether a tag describes an object visible on a picture, given a picture and a tag.**

### 3 What Does the Picture Show? An Empirical Study with Two Human Raters

In an empirical study we sampled data from Flickr and let two human raters answer the question which tags refer to visible entities on the corresponding pictures. Based on this sample data, we investigated **(a) the proportion of descriptive tags and (b) the reliability of human ratings.** The latter indicates whether the given problem, identifying descriptive tags, can reliably be decided given a picture and its tags. If the ratings differ substantially across time or different raters for instance, this would indicate that the problem can only be decided taking into account additional information such as e.g. time-of-the-day or mood of the rater.

<sup>4</sup> A villa near Berlin’s Wannsee (a lake)

<i>Given tags:</i>		
Flower	Garten	garden
Haus Liebermann	Malve	Weitwinkel
hollyhock	wide angle	Wannsee
Germany	NaturesFinest	
<i>Tags probably describing content:</i>		
Malve	hollyhock	Flower
Garten	garden	Haus Liebermann



**Fig. 1.** Only part of the tags describe the content of the picture. Other tags describe where it was taken and which camera calibration was used. Lower- and uppercase writing was taken over from the original Flickr tags. The Flickr page of this photo is online at <http://flickr.com/photos/sevenbrane/2631266076/>. The photo is licensed under the Creative Commons <http://creativecommons.org/licenses/by-nc/2.0/deed.en>

### 3.1 Setup of the Study

**Data Sets** As data source for our study we chose Flickr, which provided us with pictures and tags given both by picture owners and visitors. Flickr is an online photo sharing and management platform. Pictures on Flickr mostly show everyday objects that are also visible to the human eye and are mostly made by handheld cameras. Tags are mostly in English (see e.g. [7]).

For the analyses, a data set (Set A) was selected. In order to arrive at Set A, a set of 500 publicly available pictures rated as “most interesting” were downloaded on July 3, 2008 from Flickr, using the Flickr API<sup>5</sup>. Pictures without tags were removed from the dataset, which left 405 photos. The photos were mostly tagged by their owners, but partly also by visitors to the pictures. The thus compiled data set for our study (Set A) consists of 3862 (*picture, tag*) pairs and was then also used for the evaluation of the algorithm described in Sec. 4.3 For investigating the reliability of human rating decisions, a further data set (Set B)

<sup>5</sup> <http://www.flickr.com/services/api/>

was needed. Set B consists of 20 randomly chosen pictures from Set A. Set B hence is a subset of Set A and it comprises 189 (*picture, tag*) pairs.

**Rating Procedure** Two human raters participated in our investigation. Both Rater 1 and Rater 2 were native German speakers, although they had a good knowledge of English. In order to define the relation “this picture shows a(n) <tag>”, for each pair (*picture, tag*), a rating has to be made, whether the tag describes something that is visible on the picture. A judgement with regard to one picture and one tag therefore consists of deciding either: “yes, this tag describes an object visible on the picture” (positive decision) or “no, this tag does not describe an object visible on the picture” (negative decision). The rating can be done by human raters or a machine (algorithm).

For the user study, Rater 1 was provided with a table of all (*picture, tag*) pairs from Set A, and with all 405 pictures that built the basis for Set A. Rater 1 was instructed to rate each (*picture, tag*) pair according to the above described rating procedure. The decision was noted in a table next to the pair (*picture, tag*). This procedure was repeated for all 405 pictures, i.e. for all 3862 (*picture, tag*) pairs. The rating procedure for Rater 2 was equal to the procedure of Rater 1, but only performed on Set B.

**Quantifying the Agreement between Different Rating Sources** If two (human or machine) raters *agree*, this means that both raters decide for a (*picture, tag*) pair either “yes, the picture shows a <tag>”, or both decide “no, the picture does not show a <tag>”. If the same rater gives a judgement at two different points in time, the same terminology applies. *Disagreement* means that one of the raters decides “yes, the picture shows a <tag>” while the other rater comes to the conclusion “no, the picture does not show a <tag>”. Aggregated over a sample of tags, the extent of *agreement* and *disagreement* can be visualized by means of a contingency table or a bar chart. The percentage of agreement (or disagreement respectively) of two raters can easily be computed by looking at the ratio between the number of judgements where both raters agree and the total number of judgements. Moreover, the agreement and disagreement can be measured using a correlation coefficient, or a contingency coefficient.

For the analyses that we describe in the remainder of this article we use the  $\Phi$  coefficient (see e.g. [8]), a measure of the degree of association between two binary variables. The binary variable in our case is the yes/no decision taken with respect to a (*picture, tag*) pair. The values of the  $\Phi$  range from  $-1$  to  $+1$ . For our purposes  $\Phi = +1$  can be interpreted as *perfect agreement*, which means that the raters agree for all rating decisions. *Perfect disagreement* is indicated by  $\Phi = -1$  and means that two raters disagree for all rating decisions. A coefficient  $\Phi = 0$  means that the rating decisions of two raters are not systematically connected at all. In other words, if  $\Phi = 0$ , the decision of one rater (whether a tag refers to an object visible on a picture) cannot be predicted from the other rater’s decision.  $\Phi$  allows testing for *statistical significance*.

Naturally, an algorithm must aim towards perfect agreement. However, any al-

gorithm that shows perfect disagreement could easily be changed into one with perfect agreement by simply inverting its decisions.

### 3.2 Percentage of Tags Referring to an Object Visible on a Picture

One question to be answered in our study was how many tags on average describe visible entities on the tagged pictures. Rater 1 rated 782 (20.3%) out of 3862 tags in Set A to describe visible objects. The low number of tags which are considered to refer to something visible on a picture is remarkable. The goal of our research was to investigate the possibility of automatically distinguishing exactly those 20.3% of descriptive tags from the other 79.7% of non-descriptive tags.

### 3.3 Reliability of the Rating Decisions

The other question that we wanted to answer concerned the reliability of the rating decisions by the human raters. Two aspects of reliability were taken into account, namely retest reliability and inter-rater reliability. In this context, retest reliability refers to the stability of the judgements of Rater 1 over time and their independency of confounding variables (e.g. mood of the rater, day of the week). If a rater is not able to make reliable decisions, this is a source of error variance, and an indication against the feasibility of deriving an algorithm which is able to approximate human ratings. Inter-rater reliability means the extent to which the result is unaffected by individual rating tendencies of Rater 1. If inter-rater reliability is low, any algorithm trying to take that decision must be personalised.

**Retest Reliability** For investigating retest-reliability, Rater 1 was asked to repeat part of her ratings two weeks after she had performed the rating procedure described in 3.1 for Set A. The repeated ratings were given on Set B (189 *(picture, tag)* pairs, a subset of Set A). At the time the retest ratings of Rater 1 were given, one of the pictures was not online anymore. Therefore Rater 1 rated only 182 *(picture, tag)* pairs in the retest round. Examination of the retest reliability gave  $\Phi = 0.84$  ( $p < .01$ ). This indicates a high retest reliability. In concrete numbers, Rater 1 assessed 182 tags twice. At the second rating, she rated 9 (4.9%) out of these 182 tags differently than the first time.

**Inter-rater Reliability** In order to assess inter-rater reliability, Rater 2 was asked to judge the tags assigned to the pictures in Set B. These ratings were compared with the original ratings of Rater 1 on Set A. Overall, 189 *(picture, tag)* pairs were judged by both raters. The agreement of the two raters was  $\Phi = 0.77$  ( $p < .01$ ). This indicates also a high inter-rater reliability. In concrete numbers, Rater 1 and Rater 2 disagreed on 14 (7.4%) out of 189 tags.

**Discussion** Since both, retest reliability and inter-rater reliability are satisfactory, it can be assumed that the presentation of a picture and a tag suffices in principle to predict the judgement of a human rater on whether the tag is in a “shows”-relation with the picture.

## 4 A Tag-based Algorithm

After having found out that given a picture and its tags the decision which tags describe visible objects on the picture can be reliably made by human raters (see above), we were interested in whether even more contextual information could be taken away. Assuming an agent only sees the tags, without the picture, how well can it guess which tags will describe visible entities on the picture?

The interest of this question, besides its academic interest, is very practical: Analysing a number of tags is easier than analysing picture content, and an algorithm which automatically can identify the “shows”-relation could clearly be useful. For instance, starting from a picture showing specific objects (e.g. a house and flowers), it could immediately retrieve pictures which show similar kinds of objects. Such an algorithm could also be useful to get a rough overview of the content of image repositories. An automatic execution of this task has the advantage of performing well also on huge amounts of data, and in producing machine-readable metadata which can be further processed automatically.

### 4.1 Quantifying the Limitation of a Tag-based Algorithm

A tag-based algorithm has knowledge about the (meaning of) tags and may additionally have prior knowledge about the pictures to be expected. In order to quantify the limitations of any tag-based algorithm on Set A with respect to the ratings of Rater 1, an optimal algorithm was constructed. The algorithm is optimal in the sense that no other tag-based algorithm can exist which agrees better with Rater 1’s ratings on Set A. The optimal algorithm always decides exactly as Rater 1 did, except for tags where the human rater gives alternating decisions. In these cases, the optimal algorithm goes for the majority of decisions. If there is a draw, the optimal algorithm chooses “no”. For instance, the tag “leaves” occurred five times, and Rater 1 decided four of the times that the tag described objects visible on the picture. The optimal algorithm stays with the majority of decisions and always decides “yes” for the tag “leaves”. Going for the majority ratings was an arbitrary decision by the authors of the study. It does not influence the overall correlation, but would influence an in-depth study on false positive and false negative decisions.

The overall agreement between the optimal algorithm and Rater 1’s ratings on Set A was computed. The correlation coefficient was  $\Phi = 0.94$  ( $p < .01$ ). In concrete numbers, the optimal algorithm agreed with the human rater on 3783 (98.0%) of 3862 tags. This very high correlation indicates that in a fixed environment, e.g. within one platform, the patterns of usage are consistent enough to make an informed guess about whether a tag represents a visible entity or not. To some extent this corresponds also to making a guess about the actual meaning of a tag.

### 4.2 Algorithm Design

Given the theoretical feasibility of a well-performing tag-based algorithm, we implemented and evaluated a WordNet-based algorithm (WN-Algorithm) as proof-

of-concept. The interest to the reader lies in the, as we think, generally applicable design, and most of all in its simplicity together with its good performance.

As was already seen, a tag-based algorithm deterministically guesses, given a tag, whether it will be visible on a given (indeed on any) picture. In order to perform well, such an algorithm must have a general knowledge about the meaning of tags, i.e. which tags describe concepts that can in principle be visible on a picture, as well as some knowledge about the domain, i.e. which pictures can be expected. The second is necessary to make reasonable decisions given tags like “Africa”, which in Flickr mostly seem to mean “taken in Africa”, while in a picture database of satellite pictures it would probably mean “Africa is visible”.

The design of WN-Algorithm follows two steps. First, every tag must be disambiguated, i.e. decided which of a tag’s possibly many meanings the algorithm shall assume. In a second step, given the meaning of a tag, the algorithm must decide whether it is likely to denote a visible entity on any given picture of the picture database.

**Disambiguation** The knowledge about meanings of words, and the general frequency of their occurrence are both encoded in WordNet [9, 10], a lexical database of English. For instance, WordNet encodes that the word “wood” may mean either the material or a forest amongst other things, but that mostly the former is meant in English. Additionally, WordNet maps English words (nouns, adjectives, adverbs, verbs) to sets of synonyms, and refers to a reasonable amount of words used as tags in Flickr ([7]). For more specialised domains, a domain ontology including references from concepts to words (e.g. via labels, comments, synonyms) would be necessary.

We actually side-stepped the issue of disambiguation by mapping each tag simply to the concept it most frequently stands for. The relevant point is to understand that this procedure is a simple heuristic. Clearly this potentially affects the proposed algorithm whenever the most frequent meaning of a word according to WordNet does not correspond to the most frequent meaning of a tag in Flickr. In order to deal with very obvious such cases, we have had to introduce stop-and-go-wordlists.

**Rules** The knowledge which concepts denote entities likely to be visible and which do not was encoded into rules based on the underlying knowledge structure, i.e. WordNet. Rules were formulated as decision criteria in terms of the WordNet hierarchy. The hierarchy was traveled from top to down, and at appropriate levels rules specified that all concepts<sup>6</sup> and instances<sup>7</sup> in the subtree below would be decided to denote visible entities on any given picture or not. The WordNet-based rules were preceded by a preprocessing stage in which stemming (stripping tags off eventual suffixes) and exclusion of adjectives, adverbs

---

<sup>6</sup> A concept refers to an idea of something. A concept often refers to something abstract, e.g. “love” or to a group of real world entities, e.g. “flower”.

<sup>7</sup> An instance refers to a specific entity in the real world, e.g. “Big Ben” is an instance of the concept “clock tower” and refers to a specific entity.

and verbs<sup>8</sup> was performed.

An exemplary rule in WN-Algorithm is “If a tag corresponds to an instance of a concept in WordNet, in most cases it does not describe a visible object on a picture in Flickr except if it is an instance of the concept ‘artefact’ ”. Consequently, for any tag which corresponds to an instance of a concept other than “artefact”, e.g. “vienna” or “mozart”, WN-algorithm guesses that the tag does not describe a visible entity on any picture in Flickr. Indeed, the tag “vienna” mostly means that the picture was taken in Vienna but not that the picture shows (the whole of) Vienna. A search for pictures tagged with “mozart” returns a lot of pictures taken in Salzburg and some from opera performances evidently of one of Mozart’s plays, but among the first 100 hits, there is not a single (!) picture showing Mozart himself. Instances of the concept “artefact” are for example “eiffel tower” or “golden gate bridge”.

### 4.3 Algorithm Evaluation

WN-Algorithm’s ratings were compared with Rater 1’s ratings on Set A. An overview of the results is given in Table 4.3. The agreement between the implemented algorithm and the Test Set is  $\Phi = 0.55$  ( $p < .01$ ). Out of 3862 (*picture, tag*) pairs, Algorithm 95 agreed with the Test Set on 3290 pairs (84.9%) and disagreed on 572 pairs.

Set into the context of the baselines given by the inter-rater reliability and the optimal algorithm, the good performance can be easily recognised: Where Algorithm 95 and the ~~Test Set~~ agree in 84.9%, the two raters agree in 92.6%. The difference to the optimal algorithm, which achieves an agreement of 98.0%, is slightly higher, which shows in our opinion the potential for personalisation. A personalised algorithm would have to take mindset, perception and personal basic level into account (see Section 5 for a detailed description).

In order to obtain indications on the generalisability of these results, the WordNet categories used by the WN-Algorithm were mapped to the WordNet categories used by Sigurbjörnsson and von Zwol in [7] on a snapshot of the Flickr database of approximately 52 million photos. A comparison shows that the proportions of WordNet categories in the two studies are similar. This indicates that our data set is representative concerning the distribution of tags over the underlying knowledge structure WordNet, and our results can be generalised to other representative data sets taken from Flickr.

## 5 Why Is It Difficult to Decide What a Picture Shows?

The example in Fig. 1 illustrates that deciding which tags are descriptive of a picture’s content is more difficult than it seems to be at first glance. The algorithm described above, simply determines that “Figure 1 shows a flower, a garden

---

<sup>8</sup> Adjectives denote qualities of possibly visible things, but there is nothing like an instance of “beautiful”, and similar for adverbs and verbs.

	WN-Algorithm	Rater 2 (Inter-rater reliability)	Optimal Algorithm
Agreement with Rater 1 [%]	84.9%	92.6%	98.0%
Agreement with Rater 1 [ $\Phi$ ]	0.55 ( $p < .01$ )	0.77( $p < .01$ )	0.94( $p < .01$ )

**Table 1.** The implemented algorithm’s agreement with Rater 1’s ratings on Set A in comparison to the inter-rater reliability and the agreement of the optimal algorithm with Rater 1’s ratings on Set A.

and a hollyhock”. In addition to this, Rater 1 identified the tags “Malve” and “Garten”, the German words for hollyhock and garden respectively. The other ratings coincide. Additionally, it could be conjectured that “Haus Liebermann” is the name of the house on Fig. 1. Depending on the knowledge of the rater (human or machine), the judgement would then differ.

In a detailed manual analysis we studied the  $(picture, tag)$  pairs on which Rater 1 and Rater 2, or Rater 1 and WN-Algorithm, disagreed. For space reasons, detailed references to the underlying test data are not given. The following list of reasons for disagreement between (human and algorithmic) raters is the condensed output of the analysis that was carried out. These reasons for disagreement illustrate at the same time the inherent difficulty in distinguishing descriptive from non-descriptive tags given a  $(picture, tag)$  pair.

**Definitional Disagreement** Two raters have a different definition of what “an object” on a picture is. This is a foundational source of disagreement, since the problem we wanted an algorithm to solve was exactly to decide on which tags describe *objects* visible on a picture.

Examples of this are the tags “agriculture” or “feeling stitchy” (which was a writing embroidered on linen on the picture in question), on which the two human raters disagreed.

**Difference in Perception** An object might have been overlooked, such as a spider web<sup>9</sup>.

**Difference in Knowledge** This may happen with technical terms, uncommon English terms or terms in an unknown foreign language. Examples are the terms “mangrove” , “hydrangea” or “jetty” . This error can also be made by algorithmic raters, e.g. if the used background knowledge is not specific enough or does not contain terms of a specific language.

We observed that in cases of lack of knowledge or lack of confidence to agree to a specific term, human raters tended to rate a tag as not describing an object shown on the picture. This might be an argument in later stages for emphasizing the minimisation of false positive rating decisions.

<sup>9</sup> The picture shows a flower with a butterfly, and a barely visible spider web:<http://flickr.com/photos/18718027@N00/2631263572>

**Difference in Mindset** Due to different mindset (educational background, culture, personal opinion, etc.), the interpretation of a tag can differ between raters. A rater might disagree with some tag, such as calling a ferret a pet<sup>10</sup>. Especially with artistic pictures, which often show surreal or imaginary scenes, also the interpretation of a picture can differ between raters up to the point where it becomes impossible to agree on specific objects that are shown on a picture.

**Disambiguation** When a tag can have multiple meanings, a rater needs to disambiguate first before a decision can be made. Raters may disagree over the meaning they assign to a tag.

In general human raters use picture context as well as a vast amount of background knowledge to disambiguate whereas an algorithmic rater disambiguates according to strict rules and only limited background knowledge.

**Algorithmic Limitations** Finally there are erroneous decisions because of inherent limitations of the used algorithm. An algorithm that is based on some structured background knowledge is also dependent on this structure. Where the structure cannot make a difference, the algorithm cannot either (except in precise case-by-case rules / stop- and go-wordlists).

In our work, for instance the central limitation is that the algorithms are tag-based, i.e. they decide for one tag regardless of the picture. The algorithm therefore disagrees with a rater who takes the picture content into account when a tag in principle describes a visible object, such as “hotel”, but the corresponding picture does not show a hotel <sup>11</sup>.

## 6 Conclusion

What does this picture show? This question is relevant for automated processing of large repositories of tagged pictures. A simple application would be tag-based retrieval of pictures showing similar content for instance.

We investigated the question of deciding which of a picture’s tags describe visible entities depicted on it from multiple angles. First, we studied the proportion of such descriptive tags over non-descriptive tags. On our test data it has turned out that only approximately 20% of all given tags relate to the directly visible picture content. We further explored the problems involved with answering such a question for human raters, and investigated retest and inter-rater reliability. A qualitative analysis led to the insight that identifying descriptive tags reliably is not a clear-cut task even for humans, because of differences in definition, knowledge, mindset, perception and disambiguation. With the goal of devising an algorithm which is able to reliably identify the “shows”-relation between a tag and a picture, we further investigated the feasibility of tag-based algorithms, which should base their decisions solely on the tags. Given a positive result, namely that within a given platform the usage pattern of one tag

<sup>10</sup> The picture shows a ferret: <http://flickr.com/photos/77651361@N00/2631585847>

<sup>11</sup> This picture shows the ocean, a piece of beach and a bird but not a hotel: <http://flickr.com/photos/26079103@N00/2630745505>.

is consistent enough to allow for tag-based decisions, we implemented a proof-of-concept solution which agrees with a human rater in  $\approx 85\%$  cases. We argue that this result is satisfactory compared with the results from retest reliability (95, 1%) and inter-rater reliability (92, 6%). Furthermore, generalisability of our results can be assumed on the one hand because of the high inter-rater reliability between the two raters, and on the other hand because of the similar tag distribution with respect to WordNet categories of Set A when compared to a much larger dataset described in [7].

Future work includes the investigation of alternative knowledge structures on which a tag-based algorithm can rely, such as e.g. OpenCyc [11]. Another open question is whether it makes sense to personalise such an algorithm and how personalisation could be incorporated into the general algorithm design. In order to generalise our approach to arbitrary resources, such as for instance web bookmarks, more research is necessary.

## References

1. Golder, S.A., Hubermann, B.A.: Usage patterns of collaborative tagging systems. *Journal of Information Science* **32**(2) (2006) 198–208
2. Volkmer, T., Thom, J.A., Tahaghoghi, S.M.M.: Modeling human judgment of digital imagery for multimedia retrieval. *IEEE Transactions on Multimedia* **9**(5) (2007) 967–974
3. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: *International Semantic Web Conference*. Volume 3729 of *Lecture Notes in Computer Science.*, Springer (2005) 522–536
4. Schmitz, P.: Inducing ontology from flickr tags. In: *Proceedings of the Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland (May 2006)
5. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference*, New York, NY, USA, ACM Press (2007) 103–110
6. Bechhofer, S., Carr, L., Goble, C.A., Kampa, S., Miles-Board, T.: The semantics of semantic annotation. In: *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, London, UK, Springer-Verlag (2002) 1152–1167
7. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In Huai, J., Chen, R., Hon, H.W., Liu, Y., Ma, W.Y., Tomkins, A., Zhang, X., eds.: *WWW*, ACM (2008) 327–336
8. Cohen, J., Cohen, P., West, S.G.: *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates (2003) ISBN: 0805822232.
9. Fellbaum, C.: A semantic network of english: The mother of all wordnets. *Computers and the Humanities* **32** (1998) 209–220
10. Miller, G.A.: Wordnet: a lexical database for english. *Commun. ACM* **38**(11) (1995) 39–41
11. OpenCyc: <http://www.opencyc.org/>. Last visited: Oct 31, 2008