Automatically Identifying Tag Types

Kerstin Bischoff, Claudiu S. Firan, Cristina Kadar, Wolfgang Nejdl, and Raluca Paiu

L3S Research Center / Leibniz Universität Hannover
Appelstrasse 9a
30167 Hannover, Germany
{bischoff,firan,nejdl,paiu}@L3S.de, cristina.kadar@gmail.com

Abstract. Web 2.0 applications such as *Delicious*, *Flickr* or *Last.fm* have recently become extremely popular and as a result, a large amount of semantically rich metadata produced by users becomes available and exploitable. Tag information can be used for many purposes (*e.g.* user profiling, recommendations, clustering *etc.*), though the benefit of tags for search is by far the most discussed usage. Tag types differ largely across systems and previous studies showed that, while some tag type categories might be useful for some particular users when searching, they may not bring any benefit to others. The present paper proposes an approach which utilizes rule-based as well as model-based methods, in order to automatically identify exactly these different types of tags. We compare the automatic tag classification produced by our algorithms against a ground truth data set, consisting of manual tag type assignments produced by human raters. Experimental results show that our methods can identify tag types with high accuracy, thus enabling further improvement of systems making use of social tags.

Keywords: collaborative tagging, classification, tag types, social media.

1 Introduction

Collaborative tagging as a flexible means for information organization and sharing has become highly popular in recent years. By assigning freely selectable words to bookmarked Web pages (*Delicious*), to music (*Last.fm*) or pictures (*Flickr*) users generate a huge amount of semantically rich metadata. Consequently, several well known tagging systems have been acquired by search engine companies to exploit this additional information during search. Especially for multimedia resources, accurate annotations are extremely useful, as these additional textual descriptions can be used to support multimedia retrieval. Prior studies, which started to investigate users' motivations for tagging and the resulting nature of such user provided annotations, discovered that both motivations for tagging, as well as the types of assigned tags differ quite a lot across systems. However, not all tags are equally useful for search. For example, a user might tag a picture on *Flickr* with some of the things depicted on it, like "flowers", "sun", "nature", or with the associated location ("London") and time ("2008"). Since such tags are factual in nature, *i.e.* they are verifiable at least by common sense, they are potentially relevant to all other users searching for pictures *e.g.* from this location. However, to provide some more context for sharing her images with friends, she may also add more subjective, contextual tags like "awesome" or "post-graduate trip", or she may refer to

R. Huang et al. (Eds.): ADMA 2009, LNAI 5678, pp. 31–42, 2009.

herself by using the annotation "my friends". Assuming a certain amount of interpersonal agreement, subjective tags may still be useful for some users. For the majority of users, the tag "awesome" for example, may be an indicator of the quality of the picture, but not for people disagreeing with popular opinion. Self reference tags on the other hand are so highly personal that another person may not understand the tag at all or associate something different with it (*e.g.* her own post-graduate trip to Asia). Thus, personal tags are not applicable to other users of the system, except from the user herself and maybe some of her friends. Still, for estimating similarity between resources or users search engines and recommendation algorithms exploiting user generated annotations but not differentiating types of tags and their interpersonal value incorporate all (frequent) tags and thus introduce noise. Being able to distinguish between the types of tags associated to resources would thus be highly beneficial for search engines and recommendation algorithms to best support users in their information needs. Besides, tag classes enable building enhanced navigation tools. While currently the user faces a potentially infinite, unordered tag space, tag classes would allow for browsing pictures, web sites or music by the different informational facets of the associated tags.

In this paper we tackle exactly this aspect presenting an approach to automatically identify tag types. We rely on a tag type taxonomy introduced in [1] and analyzed with respect to the potential of the tag classes for search. Our approach is applicable to any tagging system and is not bound to one particular resource type.

2 Related Work

Recent scientific work has started examining tagging behaviors, tag types and automatic tag classification, many studies focusing on one specific collaborative tagging system.

2.1 Tagging Motivations and Types of Tags

Analyses of collaborative tagging systems indicate that incentives for tagging are quite manifold and so are the kinds of tags used. According to [2], organizational motivations for enhanced information access and sharing are predominant, though also social motivations can be encountered, such as opinion expression, attraction of attention, self-presentation [2,3]. Which of those incentives is most characteristic for a particular system seems to vary, depending on tagging rights, tagging support, aggregation model, *etc.* – all influencing why certain kinds of tags are used. [3] and [4] indicate that in free-for-all tagging systems like *Last.fm*, opinion expression, self-presentation, activism and performance tags become frequent, while in self-tagging systems like *Flickr* or *Delicious* users tag almost exclusively for their own benefit of enhanced information organization.

Despite the different motivations and behaviors, stable structures do emerge in collaborative tagging systems [3,5,6]. The evolving patterns follow a scale-free power law distribution, indicating convergence of the vocabulary to a set of very frequent words, coexisting with a long tail of rarely used terms [5,6]. Studying the evolution of tagging vocabularies in the MovieLens system, [7] use controlled experiments with varying system features to prove how such design decisions heavily influence the convergence process within a group, *i.e.* the proportions "Factual", "Subjective" and "Personal" tags

will have. According to these results, being able to display automatically identified "Factual" tags only would lead to even more factual and interpersonally useful tags. Similarly, in their paper on collaborative tag suggestions, [8] introduce a taxonomy of five classes: Content, Context, Attribute, Subjective and Organizational tags.

[1] introduce an empirically verified tag type taxonomy comprising eight categories (Topic, Time, Location, Type, Author/Owner, Opinions/Quality, Usage context, Self reference) that is applicable to any tagging system, not bound to any particular resource type. Besides establishing type distributions for *Last.fm*, *Delicious* and *Flickr*, the authors discuss the potential of the different identified categories for supporting search. A complementing query log analysis showed that *e.g.* highly personal self-reference tags are indeed not used in querying a web search engine. Similarly, subjective usage context and opinions are rarely queried for, nor judged very useful for searching public web pages. Only for music these queries play an important role with people often searching for "wedding songs" or "party music". Here, interpersonal agreement seems higher due to the restricted domain and, probably, shared culture. In the present paper we will make use of this taxonomy and focus on automatically classifying tags accordingly.

2.2 Automatic Classification of Tags

So far, there have been only few studies trying to automatically categorize user tags. However, they all focus solely on specific domains and make no statements about the generalizability of their approaches to other areas apart from the original ones. Focusing on the domain of pictures, [9] try to extract event and place semantics from tags assigned to *Flickr* photos - making use of location (geographic coordinates) and time metadata (timestamp: upload or capture time) associated with the pictures. The proposed approach relies on burst analysis: tags referring to event names are expected to exhibit high usage patterns over short time periods (also periodically, *e.g.* "Christmas"), while tags related to locations show these patterns in the spatial dimension.

In [10], different tag categories used by users to annotate their pictures in *Flickr* are analyzed automatically. Using the WordNet lexical database the authors are able to classify 52% of their sample tags into the WordNet categories: Location (28%), Artefact/Object (28%), Person/Group (28%), Action/Event (28%), Time (28%) or Other (27%). However, tag classification is not the main focus of the paper, the authors being rather interested in recommending tags to users for supporting them in the annotation process. The authors of [11] map *Flickr* tags to anchor text in Wikipedia articles which are themselves categorized into WordNet semantic categories. Thus the semantic class can be inferred – improving the classifiable portion of *Flickr* tags by 115% (compared to using WordNet only). Given a set of *Delicious* bookmarks and tags assigned by users, [12] investigate the predictability of social tags for individual bookmarks. The proposed classification algorithms make use of the page's textual content, anchor text, surrounding hosts, as well as other tags already applied to the URL. This way, most tags seem to be easily predictable, page text providing the superior attributes for classification.

Thus, existing approaches often focus on predicting certain tag types only and they do so within one particular tagging system. Some techniques are restricted to the system used, *e.g.* as they require additional metadata [9] or assume content to be textual [12]. In contrast to previous work, we present a general approach to tag type classification applying our algorithms on collections containing different kinds of resources.

3 Tag Type Taxonomy

For automatically classifying user generated tags according to their functions into types, we chose the tag type taxonomy presented in [1]. This scheme was build upon the classification presented by [3] and adapted to be applicable for various types of resources (music, Web pages and pictures). It is fine grained enough for distinguishing different tag functions and the associated interpersonal value of the corresponding tag types and, more important, it was tested for its reliability. Table 1 shows the eight classes with corresponding example tags, found in the three systems *Last.fm*, *Delicious* and *Flickr*.

Table 1. Tag type taxonomy with examples of the used tagging systems (from [1])

Nr	Category	***Delicious***	***Last.fm***	***Flickr***
1	Topic	*webdesign, linux*	*love, revolution*	*people, flowers*
2	Time	*daily, current*	*80s, baroque*	*2005, july*
3	Location	*slovakia, newcastle*	*england, african*	*toronto, kingscross*
4	Type	*movies,mp3*	*pop, acoustic*	*portrait, 50mm*
5	Author/Owner	*wired, alanmoore*	*the beatles, wax trax*	*wright*
6	Opinions/Qualities	*annoying, funny*	*great lyrics, yum*	*scary, bright*
7	Usage context	*review.later,travelling*	*workout, study*	*vacation, honeymoon*
8	Self reference	*wishlist, mymemo*	*albums i own, seen live*	*me, 100views*

Topic is probably the most obvious way to describe an arbitrary resource, as it describes what a tagged item is about. For music, topic was defined to include theme (*e.g.* "love"), title and lyrics. The Topic of a picture refers to any object or person (*e.g.* "clowns") displayed, for web sites, it is associated with the title or the main subject (*e.g.* "data mining"). Tags in the **Time** category add contextual information about hour, day, month, year, season, or other time related modifiers. It may tell the time when a picture was taken (*e.g.* "2004"), a song was recorded (*e.g.* "80s"), a Web page was written or its subject event took place (*e.g.* "November 4"). Similarly, **Location** adds additional information, telling us about the country or the city, elements of the natural landscape, sights, nationality or place of origin. It can be the place where a concert took place (*e.g.* "Woodstock"), where a picture was taken (*e.g.* "San Francisco") or a location in a Web page (*e.g.* "USA"). Tags can also specify the **Type** of a resource – *i.e.* what something is. In general it refers to types of files (*e.g.* "pdf"), media (*e.g.* "movie") or Web pages (*e.g.* "blog"). For music this category contains tags specifying the music genre (*e.g.* "hip-hop"), as well as instrumentation (*e.g.* "piano"). For images it includes photo types (*e.g.* "portrait") as well as photographic techniques (*e.g.* "macro"). Yet another way to organize resources is by identifying the **Author/Owner** of a resource. It specifies who created the resource: the singer or the band name for songs, the name of a blogger or a photographer. It can also refer to the owner of the resource: a music label, a news agency or a company. Other tags like "depressing", "funny" or "sexy" contain subjective comments on the characteristics or on the quality of a resource. Users make use of such **Opinion** tags to express opinions either for social motivations or just for simplifying personal retrieval. **Usage context** tags suggest what to use a resource for,

the context in which the resource was collected or the task for which it is used. These tags, although subjective, may still be a good basis for recommendations to other users. They can refer for example to a piece of music suitable to "wake up", a text "toRead" or a URL useful for "jobsearch". Last, the **Self reference** category contains highly personal tags, only meaningful and helpful for the tagger herself. Typical examples are "my favorite song", "home" *e.g.* referring to the start page of a site or "my friend" to indicate the presence of the user's friend on some picture.

Resources usually have more than one single tag associated with them, the tags falling into several categories. Our methodology focuses on automatically classifying all these tags from three different data sets into the eight functional categories.

4 Data Sets

For our experiments we used data from some well known collaborative tagging systems covering three different types of resources: music files, general Web pages, and pictures. For the later experiments the raw tags are used, *i.e.* no lemmatizing is applied.

We used an extensive subset of *Last.fm* pages corresponding to tags, music tracks and user profiles fetched in May 2007. We obtained information about a total number of 317,058 tracks and their associated attributes, including track and artist name, as well as tags for these tracks plus their corresponding usage frequencies. Starting from the most popular tags, we found a number of 21,177 different tags, which are used on *Last.fm* for tagging tracks, artists or albums. For each tag we extracted the number of times each tag has been used as well as the number of users who used the tag.

The *Delicious* data for our analysis was kindly provided by research partners. This data was collected in July 2005 by gathering a first set of nearly 6,900 users and 700 tags from the start page of *Delicious*. These were used to download more data in a recursive manner. Additional users and resources were collected by monitoring the *Delicious* start page. A list of several thousand usernames was collected and used for accessing the first 10,000 resources each user had tagged. The resulting collection comprises 323,294 unique tags associated with 2,507,688 bookmarks.

For analyzing *Flickr* tags, we took advantage of data crawled by our research partners during January 2004 and December 2005. The crawling was done by starting with some initial tags from the most popular ones and then expanding the crawl based on these tags. We used a small portion of the first 100,000 pictures crawled, associated with 32,378 unique tags assigned with different frequencies.

5 Automatic Tag Type Classification

For automatically classifying *Last.fm*, *Flickr* and *Delicious* tags, we propose two basic approaches to categorization, depending on the category that needs to be identified. Though some tags could be assigned to more than one functional type, we will categorize each tag according to its most popular type, mainly to make evaluation of automatic classification, *i.e.* human assessment of a ground truth data set feasible. For this, we use both straight forward matching rules against regular expressions and table look-ups in predefined lists, as well as more complex model-based machine learning algorithms.

5.1 Rule-Based Methods

Five of the eight tag type categories can be identified by using simple rules, implemented as regular expressions, or table look-ups in predefined lists.

Time. Spotting time-tags is done with the help of both several date/time regular expressions and by using lists of weekdays, seasons, holiday names, *etc.* The same predefined lists where used for all three systems. This approach can easily capture most time tags – since time vocabulary of the predominately English tags is rather restricted. Less trivial approaches, like detecting time related tags as bursts over short time periods [9], on the other hand, require time related metadata (*e.g.* upload) that is not present in all tagging systems. In total we used 19 complex regular expressions containing also 106 predefined time-related expressions (*e.g.* "May", "Thanksgiving", "monthly").

Location. For identifying location tags in *Last.fm*, *Flickr* and *Delicious*, we made use of the extensive knowledge provided by available geographic thesauri. From GATE[1], an open source tool for Natural Language Processing, a total of 31,903 unique English, German, French and Romanian location related words were gathered. These terms comprised various types of locations: countries (with abbreviations), cities, sites, regions, oceans, rivers, airports, mountains, *etc.* For *Delicious*, the list needed to be slightly adapted by manually excluding some (about 120) extremely common words (*e.g.* "java", "nice", "church") in order to assure better accuracy[2].

Type. Since the Type category, denoting what kind of resource is tagged, is system/resource dependent, separate lists were used for the three systems. A list of 851 music genres gathered from AllMusic[3] was used in order to identify type tags in the *Last.fm* data set. This inventory of genres is highly popular and also used in ID3 tags of MP3 files. As music is only a (small) part of resources tagged in *Delicious*, we gathered a list of 83 English and German general media and file format terms, *e.g. document, pdf, foto, mpg* or *blog, messenger*. For *Flickr*, the type or genre list (45 items) covers besides file formats also picture types (like *portrait* or *panoramic*), photographic techniques (like *close-up* or *macro*) or camera-related words (like *megapixel, shutter*).

Author/Owner. From the information available on *Last.fm*, *i.e.* the tracks collected, a huge catalogue of artist names resulted, against which candidate tags were matched to identify whether a tag names the author or owner of a resource. In case of *Delicious* with its wide variety of Web pages bookmarked, finding the author or owner is not trivial. Since processing of a page's content and possibly extraction of named entities seems a costly procedure, we made use of an inexpensive heuristic assuming domain owners/authors to appear in a Web page's URL. With the help of regular expressions, we checked whether the potential owner or the author of the resource appears inside the corresponding URL (`http://xyz.author.com`). For *Flickr*, classifying tags into the Author/Owner category was not possible, as pictures are mostly personal and no user-related information was included in our data set.

[1] `http://gate.ac.uk/`

[2] Can be automated *e.g.* by filtering words whose most popular WordNet synset is not a location.

[3] `http://www.allmusic.com/`

Self reference. For identifying self reference tags from *Last.fm*, we created an initial list of 28 keywords, containing references to the tagger herself in different languages (*e.g.* "my", "ich" or "mia") and her preferences (*e.g.* "favo(u)rites", "love it", "listened"). For *Delicious* we adapted the list slightly to include also structural elements of a Web site (like "homepage", "login" or "sonstiges") that do not appear in the music tagging portal. Finally, for *Flickr* the list was adapted to include some personal background references, like "home" or "friends".

The rule-based methods are run over all tags to be classified; the remaining, unclassified tags are then used for training the classifiers. This filtering simplifies the subsequent task of learning to discriminate topic, opinion and usage context tags.

5.2 Model-Based Methods

Since building a reasonably comprehensive register of topics, usage contexts or opinion expressions is due to practically inexhaustible lists impossible, model-based machine learning techniques are necessary for identifying these kinds of tags. To find the Topic, Usage context and Opinion tags, different binary classifiers were trained to decide, based on given tag features, whether a tag belongs to the respective tag class or not. Here, we used classifiers available in the machine learning library Weka[4].

Classification Features. For all three systems *Last.fm*, *Delicious* and *Flickr*, we extracted the same features to be fed into the binary classifiers: Number of users or tag frequency respectively, Number of words, Number of characters, Part of speech, and Semantic category membership.

Number of users is an external attribute directly associated to each tag, measuring prevalence in the tagging community, and thus indicating a tag's popularity, relevance and saliency. For *Flickr*, we used the absolute usage *frequency* since our data does not contain the necessary user-tag tuples and it can be considered to be an equally useful, though different, indicator of popularity. Since it has been suggested that, often highly subjective opinion tags in *Last.fm*– like "lesser known yet streamable artists" – exhibit both a higher *number of words* and *number of characters* [4], we used these intrinsic tag features as well for training our classifiers. Similarly, many of these opinion tags are adjectives while topic tags are mostly nouns [3]. Thus, we included *part of speech* as additional feature. For determining word class, we employ the lexical database WordNet 2.1[5]. In form of a derived tree of hypernyms for a given word sense, WordNet also provides valuable information about the semantics of a tag. The three top level categories extracted from here complete our tag feature set. For *Last.fm* with its multi-word tags, we collected the latter two features for each word in the tag, *i.e.* we matched all terms individually if the phrase as a whole did not have a WordNet entry.

Sense Disambiguation and Substitution. For exploiting tag information like part of speech and WordNet category during machine learning, choosing the right meaning of a tag, for example "rock", is critical. Since statistical or rule-based part-of-speech

[4] `http://www.cs.waikato.ac.nz/ml/weka/`
[5] `http://wordnet.princeton.edu/`

tagging can not be applied for the one-word tags found in *Delicious* and *Flickr*, we decided to make use of the rich semantic information provided implicitly through tag co-occurrences. For the *Last.fm* and *Delicious* sample tags, we extracted all co-occurring tags with the corresponding frequencies. To narrow down potential relations, we computed second order co-occurrence. For all sample tags, we determined similarity with all other tags by calculating pairwise the cosine similarity over vectors of their top 1000 co-occurring tags. A very high similarity should indicate that two tags are almost synonymous because they are so frequently used in the same context (*i.e.* co-tags) – the two tags themselves rarely appearing together directly [13]. Given an ambiguous tag, we now search for the newly identified similar tags in the definitions, examples and synset words in WordNet. If this does not decide for one meaning, then by default the sense returned by WordNet as most popular is chosen. Since some tags are not found in WordNet at all, we make further use of similar tags by taking the most similar one having a match in WordNet as a substitute for the original tag. Due to missing coccurrence relationships for *Flickr*, neither disambiguation nor substitution could be applied.

To build models from the features listed that enable finding Topic, Usage context and Opinion tags from our sample tags, we experimented with various machine learning algorithms Weka offers: Naïve Bayes, Support Vector Machines, C4.5 Decision Trees, *etc.* For each, we moreover used different combinations of the basic features described. As the Weka J48 implementation of C4.5 yielded the best results, only the results obtained with this classifier are presented in the following section on evaluation results.

6 Results and Discussion

6.1 Ground Truth and Evaluation

For evaluating the proposed algorithms, we built a ground truth set containing sample tags from each system that were manually classified into one of the eight categories. To make manual tag categorization feasible a subset of 700 tags per system was assessed. Thus, we intellectually analyzed 2,100 tags in total. The samples per system included the top 300 tags, 200 tags starting from 70% of probability density, and 200 tags beginning from 90% – prior work suggests that different parts of the power law curve exhibit distinct patterns [5]. Clearly, such classification schemes only represent one possible way of categorizing things. Quite a few tags are ambiguous due to homonymy, or depending on the intended usage for a particular resource they can fall into more than one category. We based our decision on the most popular resource(s) tagged. On a subset of 225 tags we achieved a good and substantial inter-rater reliability for this scheme and method – a Cohen's Kappa value of κ 0.7. In general, it was often necessary to check co-occurring tags and associated resources to clarify tag meaning (see also [1]).

For measuring the performance of our tag type classification algorithms we use classification accuracy. For the model-based methods we perform a 10-fold cross-validation on the samples, and for the rule-based method we compute the accuracy by determining the number of true/false positives/negatives. Table 2 summarizes results for all systems and classes. It shows the best performing features, the achieved accuracy, precision and recall, and the percentage of tags (*i.e.* the sample of 700 tags) belonging to a

Table 2. Best results for rule-based and model-based methods. *(Features: POS=part of speech, C=WordNet categories, F=tag frequency, N=number of words and characters, RegEx=regular expressions, List=list lookup).*

	Class	Features	Accuracy	P	R	% Man.	% Auto.
Delicious	Topic	POS,C	81.46	83.89	96.38	67.14	76.00
	Time	RegEx,List	100.00	100.00	100.00	0.86	0.86
	Loc.	List	97.71	70.37	70.37	3.86	3.86
	Type	List	93.71	66.67	42.86	8.00	5.14
	Author	RegEx	70.20	9.85	38.57	6.29	2.14
	Opinion	N,POS,C	93.40	0.00	0.00	5.14	0.00
	Usage	POS,C	89.66	0.00	0.00	7.86	0.14
	Self ref.	List	99.00	33.33	16.67	0.86	0.29
	Unknown						*11.57*
Flickr	Topic	F,POS,C	79.39	84.62	88.82	46.07	45.92
	Time	RegEx,List	98.86	93.10	81.82	4.72	4.15
	Loc.	List	86.70	76.88	72.68	26.18	21.89
	Type	List	95.99	84.62	29.73	5.29	1.72
	Author	N/A				0.14	
	Opinion	N,POS,C	93.21	82.86	55.77	7.44	5.87
	Usage	N,POS,C	85.48	39.53	32.08	7.58	4.58
	Self ref.	List	97.85	100.00	16.67	2.58	0.43
	Unknown						*15.45*
Last.fm	Topic	F,N	90.32	0.00	0.00	2.43	0.00
	Time	RegEx,List	99.14	66.67	66.67	1.29	1.29
	Loc.	List	97.43	87.04	81.03	8.29	7.71
	Type	List	77.14	91.60	60.89	51.14	33.71
	Author	List	88.65	58.67	43.56	8.14	3.29
	Opinion	F,N,POS,C	74.73	79.84	83.06	17.71	18.43
	Usage	POS,C	79.57	62.96	37.78	6.43	5.29
	Self ref.	List	98.71	92.59	78.13	4.57	3.71
	Unknown						*26.57*

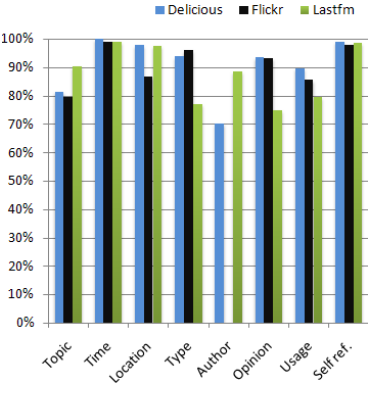

Fig. 1. Accuracy per class and system

certain category: both the real, manual value ("Man.") and the predicted, automatic value ("Auto."). A graphic representation of the accuracies is given in Figure 1.

6.2 Performance of Rule-Based Methods

The regular expressions and table look-ups performed very well in predicting the five categories Time, Location, Type, Author/Owner and Self-reference. With about 98% accuracy, performance was especially satisfactory for the highly standardized Time tags as well as for Self reference tags. However, accuracy is considerably lower for Type in *Last.fm*. This is mainly due to the used lists not being exhaustive enough. For example, the list of genres did not contain all potential sub-genres, newly emerging mixed styles or simply spelling variants and abbreviations. Its quota decreased progressively with the "difficulty" of the data set, *i.e.* the less frequent and more idiosyncratic the tags became. Since such handcrafted lists are never complete, automatic extension of the initial set should be achieved by expanding it with similar tags, *e.g.* based on second order co-occurrence. Similarly, our artist database did not contain all naming variants for a band or a singer and it had wrong entries in the artist's rubric. Allowing for partial matching, on the other hand, adds noise and results in predicting a much larger proportion of tags to denote Author/Owner than in the ground truth. For *Delicious* similarly, the regular

expressions-based method just found a portion of the tags of interest, while more rules *e.g.* including named entity recognition would probably lead to many false positives. Last but not least, system specific design choices influence the accuracy of regular expressions for *Delicious* and *Flickr*. Since space characters in tags are not allowed here, compound names are written together (like "dead.sea", "sanfrancisco", "seattlepubliclibrary") and some location names may range over multiple tags (*e.g.* "new" and "york").

6.3 Performance of Model-Based Methods

The C4.5 decision tree yielded extremely good results for tag classification into Topic, Opinion and Usage tags. From the different intrinsic and extrinsic tag attributes used as features, part of speech and the semantic category in WordNet were present for all best performing classifiers, except for Topic in *Last.fm*. Here, number of users and number of words and characters alone achieved the best results. The number of words and characters obviously helped identifying Opinion tags in all three systems as well as Usage tags in *Flickr*. However, as a consequence of the relatively small training set of 700 tags as well as the highly unbalanced 'natural' distribution of tags over the three categories, robustness needs to be improved. Although in training the classifiers the set of positive and negative examples have been balanced, some classes had very few postive examples to learn from. For *Delicious* and *Flickr* the rate of false negatives is very high for the rare Opinion and Usage tags. Thus, none (for *Delicious*) or only part of the true Opinion and Usage tags are found. In contrast, almost all true Topic tags are correctly identified, but at the same time the number of predicted Topic tags overestimates the real proportion in the ground truth for *Delicious* and *Flickr*. The opposite happens for *Last.fm*. The classifiers learn well to reject non-Topic and non-Usage tags, but they also miss more than half of the true positives. Thus, our classifiers reinforce the tendencies to focus on one particular tag type depending on the system.

Nevertheless, the average accuracy is good, lying between 82% and 88%. As shown in Table 2 and Figure 2, except for Opinion in *Delicious* and Topic in *Last.fm*, the machine learning algorithms perform well in predicting tag type shares per system correctly. For example, the Opinion classifier matches 18.43% of the tags in *Last.fm*, compared to 17.71% by human rating.

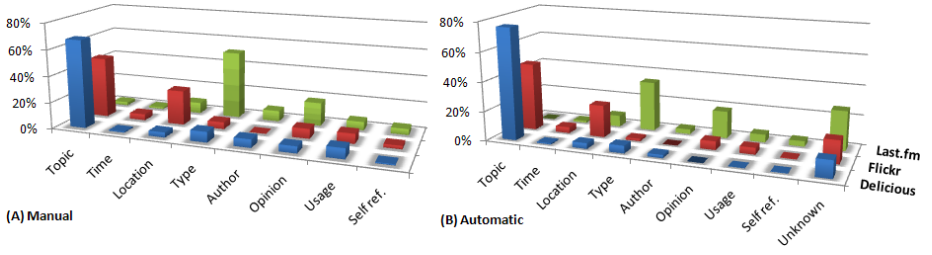

Fig. 2. Tag distribution per tagging systems: (A) manual assignment, (B) automatic assignment

6.4 Word Sense Disambiguation

Exploiting similar tags, extracted by computing second order co-occurrence, during learning improves classification performance on average by only 2% for *Last.fm*, while there is no noticeable difference for *Delicious*. Although using this method some meaningful disambiguations can be performed and a considerable part of tags not directly found in WordNet can be substituted, it does not have a big influence on classification accuracy. Some positive examples for similar tags for *Delicious* capturing synonyms, translations or simply singular/plural variations would be: "flats" and "Home.Rental", "Daily.News" and "noticias" or "technique" and "techniques". For *Last.fm*, we could find pairs like "relaxing" and "calm", "so beautiful" or "feelgood tracks". Though quite some of the similar tags found seem not to be synonymous, the strategy proved successful for disambiguation as (almost) synonymous and even strongly related words usually explain the meaning of a word. For example in the case of *Last.fm*, tags like "rock" or "pop" were correctly disambiguated and the musical meaning was chosen.

6.5 Overall Results

The linear average of all accuracies is 89.93%, while a more meaningful average, weighted by the real (*i.e.* manual) percentages of tags for each class, is 83.32%. This measure accounts for the different occurrence frequencies of the distinct tag types in the ground truth data. The weighted average values per system are: *Delicious*- 83.93%, *Flickr*- 85.07%, *Last.fm*- 81.08%. As initially shown in [1] for a smaller sample, tag class distributions vary significantly across the different systems. We observe that vocabulary and tag distribution depend on the resource domain, *e.g.* images and Web pages can refer to any topic, whereas music tracks are more restricted in content, thus leading to a more restraint and focused set of top tags. The most numerous category for *Delicious* and *Flickr* is Topic, while for *Last.fm* Type is predominant, followed by Opinion. A portion of tags could not be classified with reasonably confidence, the percentage for the "Unknown" tag type varying between 12% and 27%. Our methods overestimate the occurrences of Topic tags for *Delicious* at the expense of Opinion tags. Similary, not all Type and Author tags could be identified for *Last.fm*. Apart from this, our methods predict comparable class shares as the human raters in the overall distribution (Figure 2).

7 Conclusions and Future Work

Tag usage is rapidly increasing in community Web sites, providing potentially interesting information to improve user profiling, recommendations, clustering and especially search. It has been shown that some tag types are more useful for certain tasks than others. This paper extended previous work by building upon a verified tag classification scheme consisting of eight classes, which we use to automatically classify tags from three different tagging systems, *Last.fm*, *Delicious* and *Flickr*. We introduced two types of methods for achieving this goal – rule-based, relying on regular expressions and predefined lists, as well as model-based methods, employing machine learning techniques. Experimental results of an evaluation against a ground truth of 2,100 manually classified sample tags

show that our methods can identify tag types with 80-90% accuracy on average, thus enabling further improvement of systems exploiting social tags.

For future work, first, multi-label classification is planned to better reflect the sometimes ambiguous nature of tags. Allowing for the prediction of multiple types per tag will render unnecessary a decision mechanism for choosing the right class from those predicted by independent binary classifiers. We also like to exploit resource features like title/description for web pages, lyrics for songs, or attributes extracted by content-based methods to learn a tags' type based on the concrete resource tagged. In addition, we intend to extend the model-based methods to enable machine learning of some categories now identified by rules or look-ups.

Acknowledgments. We are greatly thankful to our partners University Koblenz/Landau and the Tagora project for providing the *Flickr* data set and from the Knowledge and Data Engineering/Bibsonomy at the University of Kassel for providing the *Delicious* data set. This work was partially supported by the PHAROS project funded by the European Commission under the 6th Framework Programme (IST Contract No. 045035).

References

1. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tags be used for search? In: CIKM 2008, pp. 193–202. ACM, New York (2008)
2. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: HYPERTEXT 2006, pp. 31–40. ACM, New York (2006)
3. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
4. Zollers, A.: Emerging motivations for tagging: Expression, performance, and activism. In: WWW Workshop on Tagging and Metadata for Social Information Organization (2007)
5. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: WWW 2007, pp. 211–220. ACM, New York (2007)
6. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
7. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: Tagging, communities, vocabulary, evolution. In: CSCW, pp. 181–190. ACM, New York (2006)
8. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: Collaborative tag suggestions. In: WWW Workshop on Collaborative Web Tagging (2006)
9. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR 2007, pp. 103–110. ACM, New York (2007)
10. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW 2008, pp. 327–336. ACM, New York (2008)
11. Overell, S., Sigurbjörnsson, B., van Zwol, R.: Classifying tags using open content resources. In: WSDM 2009, pp. 64–73. ACM, New York (2009)
12. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: SIGIR 2008, pp. 531–538. ACM, New York (2008)
13. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)