Final Project: On The Use and Abuse of Collaborative Tagging Data

Paul Heymann*
(Dated: December 15, 2006)

Many applications could benefit from labeled metadata of the sort which collaborative tagging systems produce. I looked at the data produced by one of these systems, the social bookmarking system del.icio.us, to determine whether or not the tags produced by these systems are useful externally to other applications as opposed to internally as a means of navigation. I found that due to a variety of factors, the tagging data created by users tends to be much less well distributed and has much more easily predictable information than systems like those of Luis von Ahn which are specifically engineered for the purpose of leveraging users to create metadata. However, I also found that the subjective, intrinsic information that users create may be more valuable than objective, extrinsic information that they might be forced to create because of its difficulty of prediction.

I. INTRODUCTION

Collaborative tagging systems have recently emerged as a good way to leverage large numbers of users to help organize very large, rapidly changing corpora which would be difficult to organize automatically, ranging from user contributed audio, photos, or video on a single web site to the web as a whole. Often, this works because users are working in their own self interest as they mark an object with a particular tag, and when all of these tags are aggregated together, the system can make assumptions about objects based on the aggregate activities of hundreds of thousands or even millions of users. Much recent work has looked at what sort of norms arise from collaborative tagging communities, whether a coherent taxonomy or folksonomy can be built from user contributed tags, and what can be inferred about objects and about tags based on a collaborative tagging dataset.

Meanwhile, there are many cases in which large labeled datasets organized by thousands of users would be immensely useful. One of the most obvious is search, an area where advances in the fundamental methods seem to have slowed, and emphasis now lies on how to gather increasing amounts of information about a given web page, either contributed by the page creator, a trusted source, or by users. Luis von Ahn's work [3, 6–8] has shown that thousands of volunteers, if enticed by an "entertaining" game, will be willing to label data for use in image search, vision research, or logic based on a knowledge base. A natural question is whether the data being generated by collaborative tagging systems designed for user information retrieval—in addition to Luis von Ahn's explicitly controlled systems—can also be used to learn, reason, or retrieve information about a domain.

In this paper, I describe my experiments investigating the extensibility to other tasks of one of the oldest types of collaborative tagging systems, the social bookmarking system del.icio.us [5]. I chose del.icio.us because it is one of the largest collaborative tagging systems and because the objects that it annotates, URLs, should be of immediate use to the search problem. However, in the case of many of my experiments, I believe that my observations are likely to be true of any tagging system, for the most part independent of its users and the objects that they annotate. Ultimately, it seems that the most valuable information in social bookmarking systems, and perhaps in collaborative tagging systems in general, may be the temporal and personal qualities of tags, rather than their organizational qualities.

II. DATA DESCRIPTION

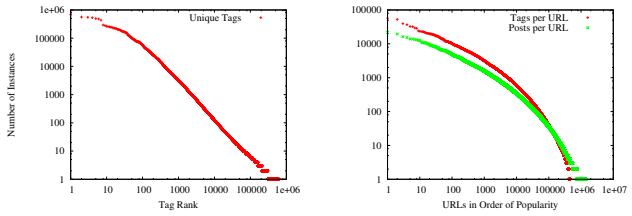
A. Preliminaries

In the course of this paper, I will use the following conventions. When a user annotates an object in a collaborative tagging system (in my case, a social bookmarking system), I will call this a *post*. A single post will consist of one or more *triples* of the form $< t_i, u_j, o_k >$ where t_i is a *tag*, u_j is the *user* making the post, and o_k is the *object* being annotated with the tag t_i, in my case a URL. When there is a triple containing tag t_i and object o_k, I say that tag t_i *annotates* object o_k. Finally, I say an object o_k is a *positive example* of a particular tag t_i if that tag annotates the object, and a *negative example* otherwise.

B. Del.icio.us Dataset

My dataset consists of $470,681$ unique URLs, $9,544,252$ posts, and $932,390$ unique tags. I gathered my data from the del.icio.us site over the course of several weeks in October 2006. I used a crawler which started at the tag "web" and expanded outwards, treating del.icio.us as a graph where each user, tag, or URL has outlinks to other users, tags, or URLs. My initial del.icio.us crawler found $1,371,941$ distinct URLs and $22,206,266$ posts of which I was able to download and

*Electronic address: heymann@cs.stanford.edu; Many thanks to Hector Garcia-Molina (my advisor) and Daniel Ramage for their advice during this project.



(a)Distribution of Tags (b)Tags/Posts Per URL

FIG. 1: The Distribution of Tags graph shows the number of tags by tag rank—in other words, each point on the X axis represents one tag, and the tags are sorted by the number of instances where someone has annotated a URL with that particular tag. The Tags Per Url and Posts Per URL graph shows the number of tags and posts by URL rank—in other words, each point on the X axis represents one URL, and the URLs are sorted by the number of instances where someone has annotated the URL with a tag or has posted the URL.

get page text for 470,681.[1] If there is any bias in my results due to my methodology for gathering data, it is most likely that it is biased towards tags or users closest to the *web* tag. There are relatively few published statistics on the size of del.icio.us—a recent article [1] put the number of posts at 53 million and the number of distinct URLs at 25 million, so my dataset is probably a substantial portion of the total posts, but probably under-represents the number of distinct URLs substantially (because these are harder to reach using my crawling method).

The number of triples per tag is distributed according to the power law distribution shown in Figure 1(a). There are 33,031,412 triples in the corpus, so each URL is annotated on average with about 70 tags. Because of the power law distribution, more than half of the tag instances in the corpus are annotations of URLs with tags that are among the top 130 tags. Likewise, the top 1000 tags on `del.icio.us` comprise over 75% of the triples in my corpus (the 1000th most common tag, *debugging*, occurs 3,153 times and is 177 times less common than the 2nd most common tag, *software*). Appendix A contains a list of the top 130 tags in my dataset.

One difference between collaborative tagging systems and systems like those created by von Ahn is that the latter meticulously control the data that users contribute. Specifically, von Ahn's systems both make it difficult to add poor metadata (often because the metadata must match another, random user who the current user cannot communicate with) and predetermine which objects any given user is allowed to annotate. By contrast, users of del.icio.us and other collaborative tagging systems can usually add whatever annotations they desire to whatever objects they desire.

This turns out to be problematic, because for an equal number of users, systems which are explicitly designed for gathering metadata like von Ahn's will produce orders of magnitude more usable metadata about the objects in the system. The reason for this is that there exist power laws over the distributions of tags to URLs and URLs to tags which are the direct result of users being able to annotate what they desire, rather than what is unlabeled in the system (see Figures 1(a) and 1(b)). The result is that if we consider our unit of work to be one post by a user, one half of the work in the system is concentrated on the top 2588 URLs, three quarters on the top 6354 and nine tenths on the top 9905, and 99 percent on the top 88830 URLs. Put another way, about 90 percent of the effort by users of del.icio.us goes in to labeling less than 2.2 percent of the data.

Arguably, most of the 90 percent of effort dedicated to the top 2.2 percent of URLs is wasted effort.[2] Furthermore, as a result of the neglect of the less popular URLs, we have much less information over which to aggregate and reason about these URLs and determine which of the metadata generated by users is "good," and which is "bad."[3] However, this is not a completely hopeless situation—it may be possible to leverage agreement with others on the most popular URLs to determine the trust or level of confidence we have in the quality of tags that a given user uses to annotate less popular URLs.[4]

III. TAGGING BIAS

I also found that the act of tagging seems to lend itself to certain types of behaviors that make the tags less valuable as organizational metadata. The primary advantage of tagging is that it can be done very quickly as a free association based on the object being tagged.[5] As a result, users can be bothered to tag and do not need any training to do so, unlike determining where in a complex hierarchy an object belongs, or what keywords out of a specific predetermined vocabulary might apply to an

[1] The unused pages were primarily either inaccessible or were not "classical" web pages, being either RSS feeds, binaries for download, or otherwise.

[2] In theory, each post could have different tags for the same URL, but in practice this is not the case.

[3] I leave these notions intentionally vague, though "bad" could mean anything from meaningless to noisy data to spam.

[4] One reason I did not pursue this task is that while this might help with determining quality of the contributions of non-adversarial users (e.g., non-spammers), without a social network or other mechanism, this metric would be relatively easy to exploit to increase one's trust according to the system.

[5] Arguably, the features that collaborative tagging systems are non-hierarchical and that their vocabularies are determined by the evolving needs of their community are very important as well, but I believe that the fundamental reason that these systems work is because the ease of tag creation overcomes whatever free rider effect there might be.

Dom. % of Tag	Tag % of Dom.	Domain
5.0%	87.7%	java.sun.com
3.2%	81.5%	onjava.com
3.1%	82.0%	javaworld.com
1.6%	67.9%	theserverside.com
1.3%	88.7%	today.java.net

TABLE I: Java tag example.

object. However, I found that in practice this free association seems to often be biased by what the user has most recently seen, and that some of the most obvious terms that users have been primed to use as tags may be determinable automatically.

A. Bias Due to Location

I found two somewhat surprising things regarding the URL of a web page and the tags that are applied to it:

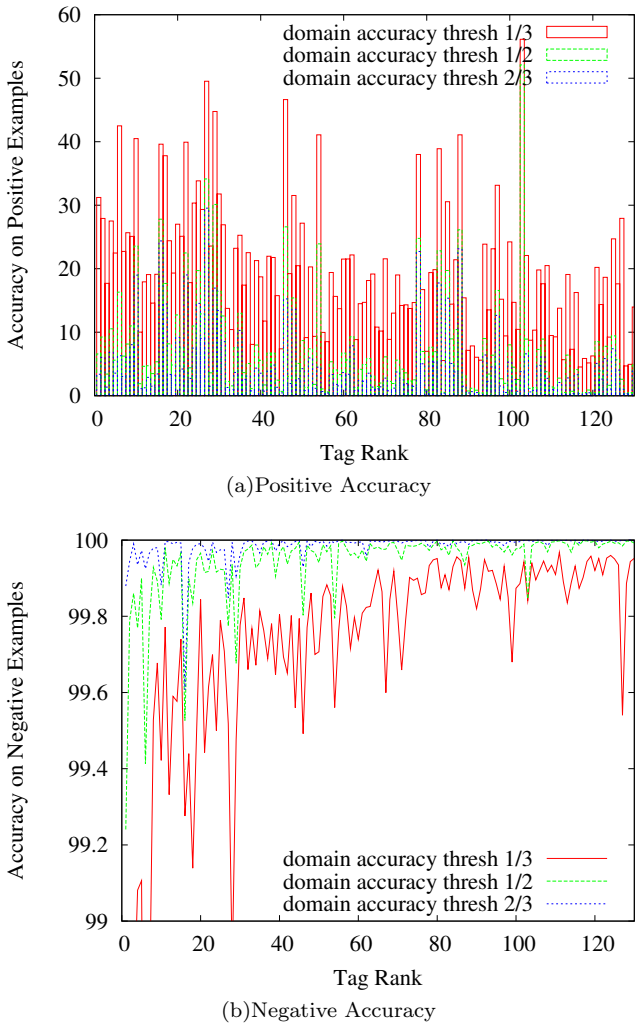
1. Users will tag an object with the location of that object, for instance, when users bookmark an interesting photograph on the web site Flickr, they will often tag it *flickr*.

2. Often, a popular site will be dedicated to a very small set of closely related topics that all fit under a single tag. For instance, for the 16th most used tag, *video*, two sites entirely dedicated to video content (youtube.com and video.google.com) make up 19.3% of the URLs which are annotated with the tag, and 60-70% of the URLs at those domains are tagged with *video*.

These two observations may be specific to social bookmarking systems, but I believe that the first might apply to any system which has "areas" or "topics" for objects, and the second might apply to any system for which there is an easy way to determine a set of tags which automatically follow from facts about the object (for instance, a user who always posts photos of cats to a collaborative tagging system for labeling photographs).

To get an idea for how prevalent these activities are, I calculate the percentage accuracy I would have for the given tag on positive or negative examples of that tag if I just classified pages as positive if they are from a domain with greater than τ_1 percent of the domain annotated with the tag, and negative otherwise. Figure 2 shows the accuracy on positive and negative examples respectively for different values of τ_1 while Table II shows the average accuracies over the top 130 tags for different values τ_1, and Table I shows an example of the phenomenon for the tag *java*.

Overall, of the top 130 tags which make up more than half of the triples in my dataset, I can recover between 5 and 20 percent of the positive examples with a false positive rate of between 3 or 4 per 1000 and 1 or 2 per



(a)Positive Accuracy

(b)Negative Accuracy

FIG. 2: Domains by Accuracy: These graphs show for the first 130 tags by rank what the accuracy on positive and negative examples is for a classifier which classifies only based on domain, and chooses positive if the domain contains more than threshold τ_1 percent URLs annotated with the given tag.

	Avg Accuracy (+)	Avg Accuracy (-)
$\tau_1 = 0.33$	19.647	99.670
$\tau_1 = 0.5$	7.372	99.943
$\tau_1 = 0.66$	4.704	99.984

TABLE II: Average accuracy predicting by domain using different values τ_1 with positive (+) and negative (-) examples.

10000 respectively. As a result, with some very conservative methods, one could probably obviate the need for a large proportion (though not the majority) of triples that users create with a relatively low false positive rate.

One interesting aspect of behavior where users add "obvious" tags that can be predicted by location is that it actually may help aid in recall for the user, but it only aids the community if the system cannot automatically

	χ^2		Mutual Information	
Tag is...	Count	% of Top 130	Count	% of Top 130
Top 1	68	52.3%	80	61.5%
Top 2	89	68.4%	98	75.3%
Top 3	93	71.5%	101	77.7%
Top 5	100	76.9%	106	81.5%
Top 10	107	82.3%	108	83.1%
Top 20	116	89.2%	113	86.9%

TABLE III: Tags vs χ^2 and Mutual Information: This shows the percentage of time and the raw counts for when a given tag in the top 130 tags is one of the top n words correlated with the tag by different measures.

Extrinsic vs Intrinsic	Organizational vs Social
What or Who Is It About	Future Retrieval
What It Is	Contribution and Sharing
Who Owns It	Attract Attention
Refining Categories	Play and Competition
Qualities/Characteristics	Self Presentation
Self Reference	Opinion Expression
Task Organizing	

TABLE IV: Tag Types: This table lists the two sets of general categories of tags, Golder and Huberman's which are based on what the tag describes, and Marlow et al.'s which identifies the purpose of the user when applying the tag.

detect pages at a given location and filter by them, and it does not aid people who would use the dataset for other purposes. While other researchers have noted that some tags (for instance, "personal") do not really add any information, I think that the use of location-based tags is an odd and remarkably frequent case where users are perhaps actually trying to explicitly help organize and annotate the corpus, but are failing to provide valuable non-obvious information.

B. Bias Due to Page Text

I found some preliminary data to suggest that users may be primed to some extent by the page text when they tag. While this is not a strong enough signal to classify just based on the tag, or a few derivatives of the tag, I did find some surprising things when I applied feature selection methods to the page text of tagged web pages.

The two feature selection heuristics I used were mutual information and χ^2 which are both commonly used in text classification. I found that very often, even in what would seem to be unlikely cases, the tag is one of the most correlated terms in page text. Table III shows the two sets of results for mutual information and χ^2. The high rank of the tags in the list of correlated terms for most tags means either or both of two things:

1. Users are primed with the page text when tagging, so they often choose tags which are in (or appear prominently in) the page text.
2. Any properly named category should have a high correlation with its contents, and users just happen to be good at choosing category names.

I believe that a little of both is probably occuring: we should expect to see the words tags represent highly correlated with themselves, but I also found evidence of tags like *interesting* which are highly correlated with themselves (according to χ^2) that one would not expect to find as a highly correlated term.

If users do in fact tend to choose words from the page text when tagging, then this may be as much of an advantage to some applications as it is a disadvantage to others. While a user who chooses a tag that occurs in the page text gives less information to someone trying to learn the general categories that the page pertains to, it may emphasize aspects of the page that the user thinks are more important.

IV. INTRINSIC AND EXTRINSIC TAGS

Various researchers have suggested different ways of grouping the fundamentally different classes of tags in tagging systems. Golder and Huberman [2] group tags by whether they are extrinsic or intrinsic to the user, and hence to what extent other users will agree that a particular object should be tagged with the tag. In contrast, Marlow et al. [4] group tags by whether their intent is organizational or social, and focus on the intent of the user when creating the tag, rather than whether the tag has meaning independent of the user. Table IV shows the two sets of types of tags.

A. Text Classification with Tags

In order to explore the predictability of tags, and hence the amount of information that different tags were adding beyond page text, I set up a series of text classification experiments on a per-tag basis. I did a set of experiments using support vector machines, and specifically Thorsten Joachim's SVMlight package, as binary classifiers for tags. For each tag, I created two sets of positive and negative examples, one training and one test set. The positive and negative examples in both sets, because I had many more negative examples than positive examples, were artificially set to be in a ratio of one positive to two negative examples. I did not modify any of the parameters of the SVMs I trained based on the test set results, so I did not do a second level development set/test set split.

B. Classifying Results

My goal in performing classification was to see which tags provided more information, or more difficult to predict information, than others. For the purposes of this paper, I avoid touching on the complex questions of what distribution of data should be trained and tested on, and what sort of false positive rate would be reasonable if one wanted to predict tags at the scale of the web. Figures 3, 4, and 5 show Best K versus Precision graphs for several example tags.

What I found was that in terms of classification, Golder and Huberman's classification of tags as extrinsic versus intrinsic[6] seems to correlate well with the relative difficulty of tag prediction of a given tag. For the most part, the relative precision of my binary classifiers for each tag fell into three categories (by decreasing precision):

1. Extrinsic tags describing topics or subject matter with very well defined vocabularies, like programming languages (*php*, *java*, *seo*) and fields or specialties (*seo*, *fonts*, *typography*, *recipes*), as well as tags which act in this way by virtue of other factors, like the common domain result discussed above (*google*).

2. Extrinsic tags describing topics or subject matter which are inherently vague, like *media*, *management*, and *tutorials* or intrinsic tags for which the community has some common standards, for example *wishlist* and *funny* (the latter of which is largely synonymous to *humor*).

3. Truly intrinsic tags which only have meaning in relation to the individual who applied them, like *interesting*, *inspiration* and *cool*.

Interestingly, the community seems to play a large part in the extent to which a particular tag will be predictable, both in general (e.g., if the community only chooses a small subset of web sites for pages tagged with *java*) and specifically when dealing with intrinsic tags. Furthermore, this means that intrinsic tags which have meaning to a particular community may be predictable, but may not transfer meanings between communities.[7]

V. CONCLUSION

I studied a large subset of the data in one of the most popular collaborative tagging systems, del.icio.us, to determine if tagging data may be useful for other applications as well as internally to improve a collaborative tagging system. What I found was neither that the data is or is not useful, but rather that collaborative tagging systems have some qualities which make them useful for some tasks and not others.

Specifically, I found that collaborative tagging systems are very inefficient for objectively labeling data as a function of effort from the users. 90% of the effort expended on tagging web pages in del.icio.us in my dataset is dedicated to the roughly two percent of pages which already have 300 or more bookmarks. Much of the labeling that is done includes tags which are either obvious from the location of the page (*flickr* or *google*) or arguably might be extracted from the page text.[8] Compared especially to systems designed for the purpose of using many users to label data, like the ESP Game, del.icio.us appears to be producing orders of magnitude less descriptive information per URL than a system which through constraints on user behavior could force the users explicitly to tag particular web pages with certain types of descriptive information.

However, on the other hand, I found that collaborative tagging systems may in fact produce a large amount of data which while not the objective labeling of objects with descriptive labels that cannot be predicted, may be more useful than that data would be. Specifically, intrinsic rather than extrinsic tags may help give opinion information not available elsewhere on a particular object, and data that results from the continuous, temporal nature of tagging systems may help provide a clue as to what is important and what topics users currently view as important.

Ultimately, it seems that the extrinsic, descriptive information in a collaborative tagging system is useful for internal navigation within the system because most users will probably be interested in the most popular items, while the intrinsic information and the information generated implicitly by users may be the most useful for external applications.

---

[6] And to a lesser extent Marlow et al.'s designations to the extent to which social versus organizational mirrors intrinsic versus extrinsic.

[7] For example, one imagines that the relatively highly predictable *wishlist* tag might not cross over from the largely technology based community of del.icio.us to the general public.

[8] My work classifying tags using SVMs has left me all the more uncertain about what a truly relevant measure of success is when attempting to automatically predict the tags on pages. Specifically, I wonder about the following issues in order of difficulty: (a) What is a reasonable distribution of positive and negative examples for a given tag to approximate the real distribution on the web in order to produce proper precision, recall, and F_1 numbers? (b) When predicting tags, do we care much more about our predictions on some pages, like the most popular or authoritative pages, than others? (c) If we try to predict tags on the web in general, will our results necessarily be flavored by the nature of the community from which we gathered our training data? Is this the case for some tags and not others?

APPENDIX A: TOP 130 TAGS

These 130 tags sorted in order by number of instances (system:unfiled occurs 682,560 times, whereas architecture occurs 40,833 times and all other tags are in between) comprise more than half of the tag instances in my del.icio.us corpus: *system:unfiled*, *software*, *reference*, *design*, *tools*, *programming*, *web*, *art*, *music*, *linux*, *news*, *howto*, *free*, *blog*, *web2.0*, *video*, *photography*, *tutorial*, *fun*, *ajax*, *webdesign*, *google*, *search*, *windows*, *javascript*, *games*, *java*, *development*, *mac*, *flash*, *css*, *cool*, *internet*, *humor*, *security*, *opensource*, *shopping*, *technology*, *science*, *books*, *tips*, *funny*, *freeware*, *business*, *tech*, *osx*, *graphics*, *php*, *politics*, *photo*, *blogs*, *travel*, *media*, *apple*, *computer*, *culture*, *photos*, *tutorials*, *education*, *hardware*, *audio*, *diy*, *research*, *history*, *productivity*, *language*, *webdev*, *online*, *social*, *html*, *toread*, *tool*, *mp3*, *writing*, *images*, *download*, *inspiration*, *python*, *rss*, *community*, *wiki*, *tv*, *photoshop*, *geek*, *ruby*, *firefox*, *movies*, *fonts*, *utilities*, *safari_export*, *interesting*, *daily*, *resources*, *maps*, *network*, *game*, *comics*, *library*, *article*, *food*, *database*, *lifehacks*, *flickr*, *email*, *code*, *hacks*, *gtd*, *xml*, *illustration*, *learning*, *magazine*, *networking*, *system:imported*, *health*, *entertainment*, *book*, *work*, *useful*, *links*, *computers*, *radio*, *unix*, *ipod*, *reviews*, *dictionary*, *visualization*, *microsoft*, *imported*, *information*, *architecture*.

[1] M. Arrington. More stats on del.icio.us, this time positive. http://www.techcrunch.com/2006/08/04/more-stats-on-delicious-this-time-positive/, Aug. 2006.

[2] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[3] L. D. Luis von Ahn. Labeling images with a computer game. In M. T. Elizabeth Dykstra-Erickson, editor, *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems*, pages 319–326. ACM Press, 2004.

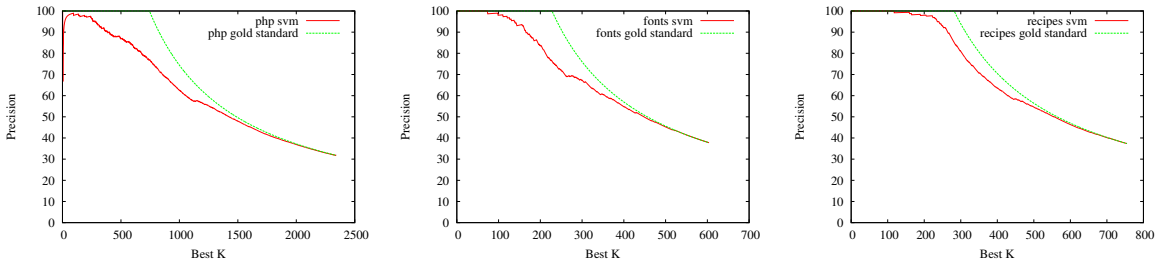
[4] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[5] J. Schachter. del.icio.us. http://del.icio.us/, Mar. 2006.

[6] L. von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

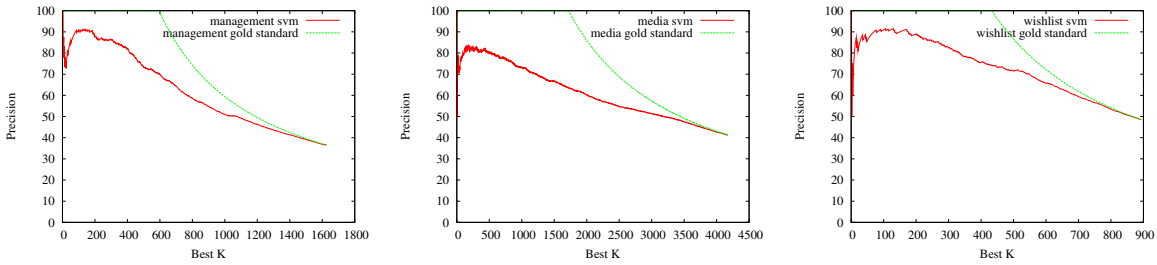
[7] L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78, New York, NY, USA, 2006. ACM Press.

[8] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64, New York, NY, USA, 2006. ACM Press.



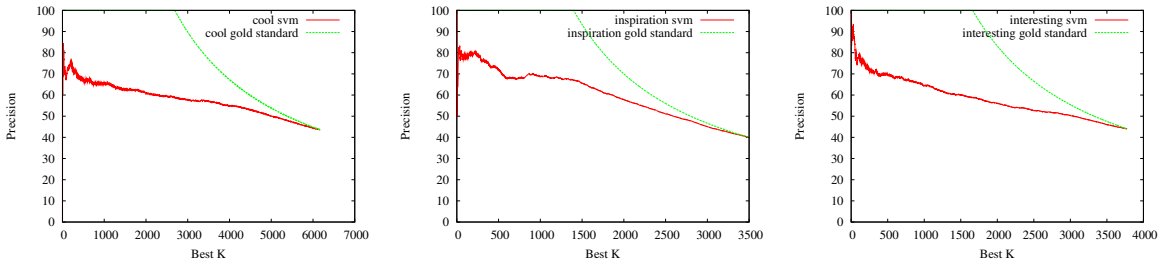
(a)PHP (b)Fonts (c)Recipes

FIG. 3: Three easy to predict tags.



(a)Management (b)Media (c)Wishlist

FIG. 4: Three tags of intermediate difficult to predict due to vagueness and other qualities.



(a)Cool (b)Inspiration (c)Interesting

FIG. 5: Three very difficult to predict intrinsic tags without very much agreed upon community meaning.