Survey on Social Tagging Techniques

Manish Gupta
University of Illinois at Urbana Champaign
gupta58@illinois.edu

Rui Li
University of Illinois at Urbana Champaign
ruili1@illinois.edu

Zhijun Yin
University of Illinois at Urbana Champaign
zyin3@illinois.edu

Jiawei Han
University of Illinois at Urbana Champaign
hanj@cs.uiuc.edu

ABSTRACT

Social tagging on online portals has become a trend now. It has emerged as one of the best ways of associating metadata with web objects. With the increase in the kinds of web objects becoming available, collaborative tagging of such objects is also developing along new dimensions. This popularity has led to a vast literature on social tagging. In this survey paper, we would like to summarize different techniques employed to study various aspects of tagging. Broadly, we would discuss about properties of tag streams, tagging models, tag semantics, generating recommendations using tags, visualizations of tags, applications of tags and problems associated with tagging usage. We would discuss topics like why people tag, what influences the choice of tags, how to model the tagging process, kinds of tags, different power laws observed in tagging domain, how tags are created, how to choose the right tags for recommendation, etc. We conclude with thoughts on future work in the area.

Keywords

Social tagging, bookmarking, tagging, social indexing, social classification, collaborative tagging, folksonomy, folk classification, ethnoclassification, distributed classification, folk taxonomy

1. INTRODUCTION

Social tagging became popular with the launch of sites like Delicious and Flickr. Since then, different social systems have been built that support tagging of a variety of resources. Given a particular web object or resource, tagging is a process where a user assigns a tag to an object. On Delicious, a user can assign tags to a particular bookmarked URL. On Flickr, users can tag photos uploaded by them or by others. Whereas Delicious allows each user to have her personal set of tags per URL, Flickr has a single set of tags for any photo. On blogging sites like Blogger, Wordpress, Livejournal, blog authors can add tags to their posts. On micro-blogging sites like Twitter, hash tags are used within the tweet text itself. On social networking sites like Facebook, Orkut, etc., users often annotate parts of the photos. Users can also provide tagging information in other forms like marking something as “Like” on Facebook. Upcoming event sites can allow users to comment on and tag events. Recently, tripletags (tags of the format namespace:key=value (e.g., geo:lat=53.1234) are becoming popular. Such a syntax can improve the usability of tags to a large extent. Using rel-tags[1], a page can indicate that the destination of that hyperlink is an author-designated tag for the current page. Rel-tags have been used by various implementation sites to tag blogs, music, links, news articles, events, listings, etc. Citation websites have tags attached to publication entries. Cataloging sites like LibraryThing and Shelfari allow users to tag books. Social news sites like Digg, SlashDot allow users to attach tags to news stories. Yelp, CitySearch and other such business/product reviews sites allow users to attach their reviews and other users to select tags to rate reviews too. Multimedia objects like podcasts, live casts, videos and music can also be tagged on sites like Youtube, imeem, Metacafe, etc. On Yahoo! Answers, you can tag an answer as positive or negative depending on how helpful it was. Tags are often used to collect such binary or multi-valued ratings or categorical decisions from users. Tags are omni-present on the web. But what led to the emergence of tagging based systems? As we shall see in this section, tags are a better way of generating metadata and prevent problems associated with fixed taxonomies in social systems.

1.1 Problems with Metadata Generation and Fixed Taxonomies

Different web portals focus on sharing of different types of objects like images, news articles, bookmarks, etc. Often to enrich the context related to these objects and thereby support more applications like search, metadata needs to be associated with these objects. However, manual metadata creation is costly in terms of time and effort [33]. Also, vocabulary of this metadata may be completely different from that of system designer or content producers or taxonomy creators or eventual users. Besides associating metadata to the objects, building a taxonomy for these social sharing systems may be useful in classifying and organizing the objects. But fixed static taxonomies are rigid, conservative, and centralized [39]. Items do not always fit exactly inside one and only one category. Hierarchical classifications are influenced by the cataloguer’s view of the world and, as a consequence, are affected by subjectivity and cultural bias. Rigid hierarchical classification schemes cannot easily keep up with an increasing and evolving corpus of items. Social systems need to hire expert cataloguers who can use same thinking and vocabulory as users and can build taxonomies which can be stable over time. Once such a hierarchy is

[1] http://microformats.org/wiki/rel-tag

created, the object creators can be asked to assign a fixed category to the object, in the hierarchy. This can induce "post activation analysis paralysis" [19] into the user. By their very nature, hierarchies tend to establish only one consistent authoritative structured vision. This implies a loss of precision, erases difference of expression, and does not take into account the variety of user needs and views.

1.2 Folksonomies as a Solution

Folksonomies and social tagging help in preventing these problems and hence provide a simpler, cheaper and a more natural way of organizing web objects. A folksonomy (folk (people) + taxis (classification) + nomos (management)) is a user-generated classification, emerging through bottom-up consensus. The term was coined by Thomas Vander Wal in the AIfIA mailing list to mean the wide-spreading practice of collaborative categorization using freely chosen keywords by a group of people cooperating spontaneously. A folksonomy can be defined as a collection of a set of users, set of tags, set of resources or objects, and a ternary relation between users, tags and resources with a time dimension [11]. Unlike formal taxonomies, folksonomies have no explicitly defined relationship between terms. All terms belong to a flat namespace, i.e., there is no hierarchy. Since users themselves tag the objects, folksonomies directly reflect the vocabulary of users [39]. Hence, a folksonomy is a simple, emergent and iterative system. It helps create the most popular way of organizing objects referred to as desire lines[2]. Apart from this, tagging provides no barriers to entry or cooperation and hence involves low cognitive cost. Tagging helps users get immediate feedback. This feedback loop leads to a form of asymmetric communication between users through metadata. The users of a system negotiate the meaning of the terms in the folksonomy, whether purposefully or not, through their individual choices of tags to describe objects for themselves. Further, folksonomies are inclusive, i.e., they include terms related to popular topics and also terms related to long tail topics. With appropriate browsing support, interlinking related tag sets is wonderful for finding things unexpectedly in a general area.

In summary, folksonomies are a trade-off between traditional structured centralized classification and no classification or metadata at all. Their advantage over traditional top-down classification is their capability of matching users' real needs and language, not their precision. Building, maintaining, and enforcing a sound controlled vocabulary is often too expensive in terms of development time and presents a steep learning curve to the user to learn the classification scheme. In other words, folksonomies are better than nothing, when traditional classification is not viable.

1.3 Outline

In this survey paper, we present a systematic detailed study of tagging literature. We first list different user motivations and different ways of tagging web objects in Section 2. There have been a lot of generative models proposed to understand the process of tagging. We present a summary of such models in Section 3. In Section 4, we present a summarization of work done on analysis of tagging distributions, identification of tag semantics, expressive power of tags versus keywords. Appropriate rendering of tags can provide useful information to users. Different visualization schemes like tag clouds have been explored to support browsing on web portals. We present some works related to such visualization studies in Section 5. When a user wishes to attach tags to an object, the system can recommend some tags to the user. A user can select one of those tags or come up with a new one. In Section 6, we discuss different ways of generating tag recommendations. In Section 7, we summarize different applications for which tags can be used. Usage of tags involves a lot of problems like sparsity, ambiguities and canonicalization. We list these problems in Section 8. Finally, we conclude with thoughts on future work in Section 9.

2. TAGS: WHY AND WHAT?

Since 2005, there have been works describing why people tag and what the tags mean. We briefly summarize such works [32; 15; 57; 6; 1; 19; 45; 33; 25; 16] below. We provide a detailed classification of user tagging motivations and also list different kinds of tags in this section.

2.1 Different User Tagging Motivations

- **Future Retrieval:** Users can tag objects aiming at ease of future retrieval of the objects by themselves or by others. Tags may also be used to incite an activity or act as reminders to oneself or others (e.g., the "to read" tag). These descriptive tags are exceptionally helpful in providing metadata about objects that have no other tags associated.

- **Contribution and Sharing:** Tags can be used to describe the resource and also to add the resource to conceptual clusters or refined categories for the value of either known or unknown audience.

- **Attract Attention:** Popular tags can be exploited to get people to look at one's own resources.

- **Play and Competition:** Tags can be based on an internal or external set of rules. In some cases, the system devises the rules such as the ESP Game's incentive to tag what others might also tag. In others, groups develop their own rules to engage in the system such as when groups seek out all items with a particular feature and tag their existence.

- **Self Presentation (Self Referential Tags):** Tags can be used to write a user's own identity into the system as a way of leaving their mark on a particular resource. E.g., the "seen live" tag in Last.FM marks an individual's identity or personal relation to the resource. Another example are tags beginning with "my" like "mystuff".

- **Opinion Expression:** Tags can convey value judgments that users wish to share with others (e.g., the "elitist" tag in Yahoo!'s Podcast system is utilized by some users to convey an opinion). Sometimes people tag simply to gain reputation in the community.

- **Task Organization:** Tags can also be used for task organization. E.g., "toread", "jobsearch", "gtd" (got to do), "todo".

[2] http://www.adaptivepath.com/blog/2009/10/27/desire-lines-the-metaphor-that-keeps-on-giving/

- **Social Signalling:** Tags can be used to communicate contextual information about the object to others.

- **Money:** Some sites like Squidoo and Amazon Mechanical Turk pay users for creating tags.

- **Technological Ease:** Some people tag because the current technology makes it easy to upload resources with tags to the web. E.g., drag-and-drop approach for attaching labels to identify people in photos. The latest photo browser commercial packages, such as Adobe Photoshop Album, adopted similar methods to support easy labeling of photos. With 'Phonetags'[3], a listener hears a song on the radio, uses her cell phone to text back to a website with tags and star ratings. Later, returning to the website, the user can type in her phone number and see the songs she had bookmarked.

2.2 Kinds of Tags

- **Content-Based Tags:** They can be used to identify the actual content of the resource. E.g., Autos, Honda Odyssey, batman, open source, Lucene.

- **Context-Based Tags:** Context-based tags provide the context of an object in which the object was created or saved, e.g., tags describing locations and time such as San Francisco, Golden Gate Bridge, and 2005-10-19.

- **Attribute Tags:** Tags that are inherent attributes of an object but may not be able to be derived from the content directly, e.g., author of a piece of content such as Jeremy's Blog and Clay Shirky. Such tags can be used to identify what or who the resource is about. Tags can also be used to identify qualities or characteristics of the resource (e.g., scary, funny, stupid, inspirational).

- **Ownership Tags:** Such tags identify who owns the resource.

- **Subjective Tags:** Tags that express user's opinion and emotion, e.g., funny or cool. They can be used to help evaluate an object recommendation (item qualities). They are basically put with a motivation of self-expression.

- **Organizational Tags:** Tags that identify personal stuff, e.g., mypaper or mywork, and tags that serve as a reminder of certain tasks such as to-read or to-review. This type of tags is usually not useful for global tag aggregation with other users' tags. These tags are intrinsically time-sensitive. They suggest an active engagement with the text, in which the user is linking the perceived subject matter with a specific task or a specific set of interests.

- **Purpose Tags:** These tags denote non-content specific functions that relate to an information seeking task of users (e.g., learn about LaTeX, get recommendations for music, translate text).

- **Factual Tags:** They identify facts about an object such as people, places, or concepts. These are the tags that most people would agree to apply to a given object. Factual tags help to describe objects and also help to find related objects. Content-based, context-based and objective, attribute tags can be considered as factual tags. Factual tags are generally useful for learning and finding tasks.

- **Personal Tags:** Such tags have an intended audience of the tag applier themselves. They are most often used to organize a user's objects (item ownership, self-reference, task organization).

- **Self-referential tags:** They are tags to resources that refer to themselves. E.g., Flickr's "sometaithurts"[4] - for "so meta it hurts" is a collection of images regarding Flickr, and people using Flickr. The earliest image is of someone discussing social software, and then subsequent users have posted screenshots of that picture within Flickr, and other similarly self-referential images.

- **Tag Bundles:** This is the tagging of tags that results in the creation of hierarchical folksonomies. Many taggers on Delicious have chosen to tag URLs with other URLs, such as the base web address for the server (e.g., a C# programming tutorial might be tagged with http://www.microsoft.com).

2.3 Categorizers Versus Describers

Taggers can be divided into two main types [27]: categorizers and describers. Categorizer users are the ones who apply tags such that the objects are easier to find later for personal use. They have their own vocabulary. Sets in Delicious is a perfect example of metadata by categorizers. On the other hand, describer users tag objects such that they are easier to be searched by others. Often tags to a single object would contain many synonyms. Vocabulary of a describer is much larger compared to an average categorizer. But a categorizer has her own limited personal vocabulary and subjective tags. ESP game is a perfect example of metadata creation by describers. Categorizers and describers can be identified using these intuitions:

- The more the number of tags that were only used once by a user, the higher the probability that the user is a describer.

- The faster the tagging vocabulary increases, the more likely it is that the person is a describer.

- A categorizer tends to achieve tag entropy that is as low as possible because he tries to "encode" her resources in a good and balanced way.

These intuitions can be formalized as metrics like tag ratio (ratio between tags and resources), orphaned tags (proportion of tags which annotate only a small amount of resources) and tag entropy (reflects the effectiveness of the encoding process of tagging).

[3] http://www.spencerkiser.com/geoPhoneTag/

[4] http://www.flickr.com/photos/tags/sometaithurts/

2.4 Linguistic Classification of Tags

Based on linguistics, tags can be classified as follows [53].

- Functional: Tags that describe the function of an object. (e.g., weapon)
- Functional collocation: These are defined by function but in addition, they have to be collected in a place (and/or time). (e.g., furniture, tableware)
- Origin collocation: Tags that describe why things are together? (e.g., garbage, contents, dishes (as in "dirty dishes" after a meal)).
- Function and origin: Tags that decribe why an object is present, what is the purpose, or where did it come from. (e.g., "Michelangelo" and "medieval" on an image of a painting by Michelangelo)
- Taxonomic: They are words that can help in classifying the object into an appropriate category. (e.g., "Animalia" or "Chordata" tag to an image of a heron)
- Adjective: They decribe the object that denotes the resource. (e.g., "red", "great", "beautiful", "funny")
- Verb: These are action words. (e.g., "explore", "todo", "jumping")
- Proper name: Most of the tags are of this category. (e.g., "New Zealand", "Manhattan bridge")

2.5 Game Based Tagging

In the ESP game, the players cannot see each other's guesses. The aim is to enter the same word as your partner in the shortest possible time. Peekaboom takes the ESP Game to the next level. Unlike the ESP Game, it is asymmetrical. To start, one user is shown an image and the other sees an empty blank space. The first user is given a word related to the image, and the aim is to communicate that word to the other player by revealing portions of the image. So if the word is "eye" and the image is a face, you reveal the eye to your partner. But the real aim here is to build a better image search engine: one that could identify individual items within an image. PhotoPlay[12] is a computer game designed to be played by three to four players around a horizontal display. The goal for each player is to build words related to any of the four photos on the display by selecting from a 7x7 grid of letter tiles. All these games, help in tagging the resources.

Problems with game-based tagging

Game-based tagging mechanisms may not provide high quality tags [28]. Maximizing your scores in the game means sacrificing a lot of valuable semantics. People tend to write very general properties of an image rather than telling about the specifics or details of the image. E.g., colors are great for matching, but often are not the most critical or valuable aspects of the image. The labels chosen by people trying to maximize their matches with an anonymous partner are not necessarily the most "robust and descriptive" labels. They are the easiest labels, the most superficial labels, the labels that maximize the speed of a match rather than the quality of the descriptor. In addition, they are words that are devoid of context or depth of knowledge. Tagging for your own retrieval is different than tagging for retrieval by people you know and even more different than tagging for retrieval in a completely uncontextualized environment.

3. TAG GENERATION MODELS

In order to describe, understand and analyze tags and tagging systems, various tag generation models have been proposed. These models study various factors that influence the generation of a tag, such as the previous tags suggested by others, users' background knowledge, content of the resources and the community influences. In this section, we present different models which have been proposed in the literature, and discuss advantages and disadvantages of these models.

3.1 Polya Urn Generation Model

Intuitively, the first factor that influences the choice of tags is the previous tag assignments. The amount of effort required to tag items may affect an individual's decision to use tags. Using suggested tags rather than one's own requires less effort. Pirolli and Card's theory of information foraging [38] suggests greater adoption of suggested tags because people adapt their behavior to optimize the information/effort ratio. Users cost-tune their archives by spending the least amount of effort needed to build up enough structure to support fast retrieval of their most useful resources. Based on this intuition, various models based on the stochastic Polya urn process have been proposed.

3.1.1 Basic Polya Urn Model

Golder and Huberman [15] propose a model based on a variation of the stochastic Polya urn model where the urn initially contains two balls, one red and one black. In each step of the simulation, a ball is selected from the urn and then it is put back together with a second ball of the same color. After a large number of draws the fraction of the balls with the same color stabilizes but the fractions converge to random limits in each run of the simulation.

This model successfully captures that previously assigned tags are more likely to be selected again. However, this basic model fails to capture that new tags will also be added into the system. So several extensions of this model have been proposed later.

3.1.2 Yule-Simon Model

Yule-Simon model [47] assumes that at each simulation step a new tag is invented and added to the tag stream with a low probability of p. This leads to a linear growth of the distinct tags with respect to time and not to the typical continuous, but declining growth. Yule-Simon model can be described as follows. At each discrete time step one appends a word to the text: with probability p the appended word is a new word, that has never occurred before, while with probability $1 - p$ the word is copied from the existing text, choosing it with a probability proportional to its current frequency of occurrence. This simple process produces frequency-rank distributions with a power law tail whose exponent is given by $a = 1 - p$.

Cattuto et al. [9] study the temporal evolution of the global vocabulary size, i.e., the number of distinct tags in the entire system, as well as the evolution of local vocabularies, that is, the growth of the number of distinct tags used in the context of a given resource or user. They find that the number N of distinct tags present in the system is $N(T) \propto T^{\gamma}$, with $\gamma < 1$. The rate at which new tags appear at time T scales as $T^{\gamma-1}$, i.e., new tags appear less and less frequently, with the invention rate of new tags monotonically decreasing very slowly towards zero. This sub-linear growth is generally referred to as Heaps' law.

3.1.3 Yule-Simon Model with Long Term Memory

Cattuto et al. [10] propose a further variation of the Simon model. It takes the order of the tags in the stream into account. Like the previous models, it simulates the imitation of previous tag assignments but instead of imitating all previous tag assignments with the same probability it introduces a kind of long-term memory. Their model can be stated as follows: the process by which users of a collaborative tagging system associate tags to resources can be regarded as the construction of a "text", built one step at a time by adding "words" (i.e., tags) to a text initially comprised of n_0 words. This process is meant to model the behavior of an effective average user in the context identified by a specific tag. At a time step t, a new word may be invented with probability p and appended to the text, while with probability $1 - p$ one word is copied from the existing text, going back in time by x steps with a probability that decays as a power law, $Q_t(x) = a(t)/(x + \tau)$. $a(t)$ is a normalization factor and τ is the characteristic time-scale over which the recently added words have comparable probabilities. Note that $Q_t(x)$ returns a power law distribution of the probabilities. This Yule-Simon model with long term memory successfully reproduces the characteristic slope of the frequency-rank distribution of co-occurrence tag streams but it fails to explain the distribution in resource tag streams as well as the decaying growth of the set of distinct tags because it leads to a linear growth.

3.1.4 Information Value Based Model

Halpin et al. [17] present a model which does not only simulate the imitation of previous tag assignments but it also selects tags based on their information value. The information value of a tag is 1 if it can be used for only selecting appropriate resources. A tag has an information value of 0 if it either leads to the selection of no or all resources in a tagging system. They empirically estimate the information value of a tag by retrieving the number of webpages that are returned by a search in Delicious with the tag. Besides of the selection based on the information value, the model also simulates the imitation of previous tag assignments using the Polya urn model. They model tag selection as a linear combination of information value and preferential attachment models. Probability of a tag x being reinforced or added can be expressed as $P(x) = \lambda \times P(I(x)) + (1 - \lambda) \times P(a) \times P(o) \times P(\frac{R(x)}{\sum R(x)})$ where λ is used to weigh the factors. $P(a)$ is the probability of a user committing a tagging action at any time t. $P(n)$ determines the number n of tags a user is likely to add at once based on the distribution of the number of tags a given user employs in a single tagging action. An old tag is reinforced with constant probability $P(o)$. If the old tag is added, it is added with a probability $\frac{R(x)}{\sum R(i)}$ where $R(x)$ is the number of times that particular previous tag x has been chosen in the past and $\sum R(i)$ is the sum of all previous tags. Overall, the proposed model leads to a plain power law distribution of the tag frequencies and to a linear growth of the set of distinct tags. It thus only partially reproduces the frequency-rank distributions in co-occurrence and resource tag streams and it is not successful in reproducing the decaying tag growth.

3.1.5 Fine-tuning by Adding More Parameters

Klaas et al. [11] present the following model: The simulation of a tag stream always starts with an empty stream. Then, in each step of the simulation, with probability I (0.6-0.9) one of the previous tag assignments is imitated. With probability BK, the user selects an appropriate tag from her background knowledge about the resource. It corresponds to selecting an appropriate natural language word from the active vocabulary of the user. Each word t has been assigned a certain probability with which it gets selected, which corresponds to the probability with which t occurs in the Web corpus. The parameter n represents the number of popular tags a user has access to. In case of simulating resource streams, n will correspond to the number of popular tags shown. (e.g., n = 7 for Delicious). In case of co-occurrence streams n will be larger because the union of the popular tags of all resources that are aggregated in the co-occurrence stream will be depicted over time. Furthermore, the parameter h can be used for restricting the number of previous tag assignments which are used for determining the n most popular tags. The probability of selecting the concrete tag t from the n tags is then proportional to how often t was used during the last h tag assignments. Using these parameters, the authors describe a model that reproduces frequency rank for both tag co-occurrence and resource streams and also simulates the tag vocabulary growth well.

3.2 Language Model

The content of resource would affect generation of tags. Hence, tagging process can also be simulated using a language model like the latent Dirichlet allocation model [5]. Tagging is a real-world experiment in the evolution of a simple language [6]. Zhou et al. [59] propose a probabilistic generative model for generation of document content as well as associated tags. This helps in simultaneous topical analysis of terms, documents and users. Their user content annotation model can be explained as follows. For document content, each observed term ω in document d is generated from the source x (each document d maps one-to-one to a source x). Then from the conditional probability distribution on x, a topic z is drawn. Given the topic z, ω is finally generated from the conditional probability distribution on the topic z. For document tags, similarly, each observed tag word ω for document d is generated by user x. Specific to this user, there is a conditional probability distribution of topics, from which a topic z is then chosen. This hidden variable of topic again finally generates ω in the tag.

3.3 Other Influence Factors

Besides above models, researchers [45; 32] have also observed that there are other factors which are likely to influence how people apply the tags.

Sen et al. [45] mention three factors that influence people's personal tendency (their preferences and beliefs) to apply tags: (1) their past tagging behaviors, (2) community influence of the tagging behavior of other members, and (3) tag selection algorithm that chooses which tags to display. New users have an initial personal tendency based on their experiences with other tagging systems, their comfort with technology, their interests and knowledge. Personal tendency evolves as people interact with the tagging system.

Experiments with Movielens dataset reveal the following. Once a user has applied three or more tags, the average cosine similarity for the nth tag application is more than 0.83. Moreover, similarity of a tag application to the user's past tags continues to rise as users add more tags. Besides reusing tag classes, users also reuse individual tags from their vocabulary. Community influence on a user's first tag is stronger for users who have seen more tags. The tag selection algorithm influences the distribution of tag classes (subjective, factual, and personal).

The community influence on tag selection in Flickr has been studied by Marlow et al. [32]. One feature of the contact network is a user's ability to easily follow the photos being uploaded by their friends. This provides a continuous awareness of the photographic activity of their Flickr contacts, and by transitivity, a constant exposure to tagging practices. Do these relationships affect the formation of tag vocabularies, or are individuals guided by other stimuli? They find that the random users are much more likely to have a smaller overlap in common tags, while contacts are more distributed, and have a higher overall mean. This result shows a relationship between social affiliation and tag vocabulary formation and use even though the photos may be of completely different subject matter. This commonality could arise from similar descriptive tags (e.g., bright, contrast, black and white, or other photo features), similar content (photos taken on the same vacation), or similar subjects (co-occurring friends and family), each suggesting different modes of diffusion.

Apart from the different aspects mentioned above, user tagging behaviors can be largely dictated by the forms of contribution allowed and the personal and social motivations for adding input to the system [32].

4. TAG ANALYSIS

To better understand social tagging data, a lot of research has been done in analyzing a variety of properties of social tagging data, such as how tags are distributed and their hidden semantics. In the following subsections we present some major analysis and results.

4.1 Tagging Distributions

Researchers began their study with analyzing tags distribution in tagging systems. They found that most of them are power law distributions, which is one of prominent features of a social tagging system.

4.1.1 Tagging System Vocabulary

As has been noted by different studies on a variety of datasets, total number of distinct tags in the system with respect to time follows a power law. However, recent studies have shown that this vocabulary growth is somewhat sublinear.

4.1.2 Resource's Tag Growth

For a single resource over time, vocabulary growth for tags also follows power law with exponent 2/3 [9]. Frequency-rank distribution of tag streams also follows a power law [11]. For some webpages tagged on Delicious, tag frequency (sorted) versus tag rank for a web page is a decreasing graph with a sudden drop between ranks 7 and 10 [11]. This may be due to an artifact of the user interface of Delicious. The graph of probability distribution of number of tags contained in a posting versus the number of tags displays an initial exponential decay with typical number of tags as 3-4 and then becomes a power law tail with exponent as high as -3.5 [9].

Researchers have also observed convergence of the tag distributions. In [17], Halpin et al. observe that majority of sites reach their peak popularity, the highest frequency of tagging in a given time period, within 10 days of being saved on Delicious (67% in the data set of Golder and Huberman [15]) though some sites are rediscovered by users (about 17% in their data set), suggesting stability in most sites but some degree of burstiness in the dynamics that could lead to a cyclical relationship to stability characteristic of chaotic systems. They also plot KL divergence between the tag frequency distributions for a resource versus the time. The curve drops very steeply. For almost all resources the curve reaches zero at about the same time. In the beginning few weeks, curve is quite steep and slowly becomes gentle as time progresses. Golder and Huberman also find that the proportion of frequencies of tags within a given site stabilizes over time.

Cattuto et al. [9] have shown the variation of the probability distribution of the vocabulary growth exponent γ for resources, as a function of their rank. The curve for the 1000 top-ranked (most bookmarked) resources closely fits a Gaussian curve at $\gamma \approx 0.71$. This indicates that highly bookmarked resources share a characteristic law of growth. On computing the distribution P(γ) for less and less popular resources, the peak shifts towards higher values of γ and the growth behavior becomes more and more linear.

Wetzker et al. [54] also show that most popular URLs disappear after peaking. They also point out that some of the tags can peak periodically, e.g., Christmas.

4.1.3 User tag vocabulary growth

There are also studies that focus on tags applied by a specific user. Golder and Huberman [15] show that certain users' sets of distinct tags grow linearly as new resources are added. But Marlow et al. [32] find that for many users, such as those with few distinct tags in the graph, distinct tag growth declines over time, indicating either agreement on the tag vocabulary, or diminishing returns on their usage. In some cases, new tags are added consistently as photos are uploaded, suggesting a supply of fresh vocabulary and constant incentive for using tags. Sometimes only a few tags are used initially with a sudden growth spurt later on, suggesting that the user either discovered tags or found new incentives for using them.

4.2 Identifying Tag Semantics

Intuitively, tags as user generated classification labels are semantically meaningful. So, research has been done for exploring the semantics of tags. These research works include three aspects: (1) Identifying similar tags, (2) mapping tags to taxonomies, and (3) extracting certain types of tags.

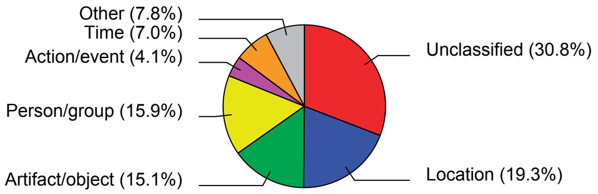
4.3 Analysis of Pairwise Relationships between Tags

In order to measure similarity of tags beyond words, researchers proposed various models to explore tags' similarity. Most of them are based on a simple assumption that tags that are similar may be used to tag the same resources, and similar resource would be tagged by similar tags. Therefore, inter tag correlation graph (tag as nodes, edges between two tags if they co-occur, weight on edge = cosine distance measure using number of times a tag was used) can be built for a tagging system. An analysis of the structural properties of such tag graphs may provide important insights into how people tag and how semantic structures emerge in distributed folksonomies. A simple approach would be measuring tags similarity based on the number of common web pages tagged by them. In section 6, we show how analysis of co-occurrence of tags can be used to generate tag recommendations.

4.3.1 *Extracting ontology from tags*

Another line of research for identifying semantics of tags is mapping tags to an existing ontology. Being able to automatically classify tags into semantic categories allows us to understand better the way users annotate media objects and to build tools for viewing and browsing the media objects. The simplest approach is based on string matching. Sigurbjörnsson et al. [46] map Flickr tags onto WordNet semantic categories using straight forward string matching between Flickr tags and WordNet lemmas. They found that 51.8% of the tags in Flickr can be assigned a semantic category using this mapping. To better assign tags to a category, content of resources associated with a given tag could be used. Overell et al. [36] designed a system to auto-classify tags using Wikipedia and Open Directory. They used structural patterns like categories and templates that can be extracted from resource metadata to classify Flickr tags. They built a classifier to classify Wikipedia articles into eleven semantic categories (act, animal, artifact, food, group, location, object, person, plant, substance and time). They map Flickr tags to Wikipedia articles using anchor texts in Wikipedia. Since they have classified Wikipedia articles, Flickr tags can be categorized using the same classification. They classify things as what, where and when. They show that by deploying ClassTag they can improve the classified portion of the Flickr vocabulary by 115%. Considering the full volume of Flickr tags, i.e., taking tag frequency into account, they show that with ClassTag nearly 70% of Flickr tags can be classified. Figure 1 shows the classification of Flickr tags.

Figure 1: Classification of Flickr tags using ClassTag system



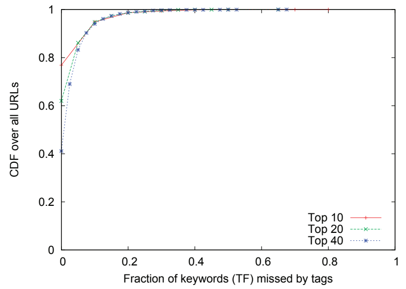
4.3.2 *Extracting place and event semantics*

Tags also contain specific information, such as locations or events. Rattenbury et al. [41] study the problem of extracting place and event semantics for Flickr tags. They analyze two methods inspired by burst-analysis techniques (popular in signal processing) and one novel method: Scale-structure Identification. The location, l_p, (consisting of latitude-longitude coordinates) associated with photo p generally marks where the photo was taken; but sometimes marks the location of the photographed object. The time, t_p, associated with photo p generally marks the photo capture time; but occasionally refers to the time the photo was uploaded to Flickr. They aim to determine, for each tag in the dataset, whether the tag represents an event (or place). The intuition behind the various methods they present is that an event (or place) refers to a specific segment of time (or region in space). The number of usage occurrences for an event tag should be much higher in a small segment of time than the number of usage occurrences of that tag outside the segment. The scale of the segment is one factor that these methods must address; the other factor is calculating whether the number of usage occurrences within the segment is significantly different from the number outside the segment. The Scale-structure Identification method performs a significance test that depends on multiple scales simultaneously and does not rely on apriori defined time segments. The key intuition is: if tag x is an event then the points in T_x, the time usage distribution, should appear as a single cluster at many scales. Interesting clusters are the ones with low entropy. For place identification, L_x is used rather than T_x. Periodic events have strong clusters, at multiple scales, that are evenly spaced apart in time. Practically, because tags occur in bursts, a periodic tag should exhibit at least three strong clusters (to rule out tags that just happened to occur in two strong temporal clusters but are not truly periodic). Overall, their approach has a high precision however a large proportion of tags remain unclassified.

4.4 Tags Versus Keywords

To identify the potential of tags in being helpful for search, there have been works that compare tags with keywords. As shown in Figure 2, given a web document, the "most important" words (both wrt tf as well as $tf \times idf$) of the document are generally covered by the vocabulary of user-generated tags [30]. This means that the set of user-generated tags has the comparable expression capability as the plain English words for web documents. Li et al. [30] found that most of the missed keywords are misspelled words or words invented by users, and usually cannot be found in dictionary.

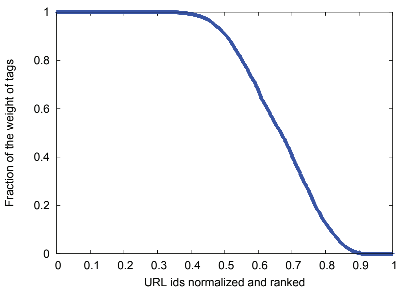
Figure 2: Tag coverage for important keywords



Further, they define tag match ratio $e(T, U)$ for tag set T associated with a URL U as ratio of weights of the tags of a particular URL that can be matched by the document. $e(T, U) = \frac{\sum_{k|t_k \in U} w(t_k)}{\sum_i w(t_i)}$. Here, $w(t)$ is the weight of tag t, i.e., the frequency of tag t in the data set. The tag match ratio represents the ratio of important tags of a URL matched by the document. Figure 3 shows the distribution of tag match ratio for URLs in their Delicious dataset.

Besides this, tags often have much more expressive power. E.g., consider the Google home page. It does not mention

Figure 3: Distribution of the tag match ratio



the phrase "search engine" anywhere. But "searchengine" can be found as a tag against the bookmarked URL http://www.google.com/ig on Delicious.

5. VISUALIZATION OF TAGS

Social tagging is one of the most important forms of user generated content. Appropriate rendering of tags can provide useful information to users. Tag clouds have been explored to support browsing on web portals, and various tag selection methods for tag clouds have been developed. Much work has been done to identify hierarchy of tags in the tag cloud construction. Visualization schemes stress on display formats and evolutionary aspect of tag clouds, etc. We review these in this section.

5.1 Tag clouds for Browsing/Search

Tag cloud, a visual depiction of user-generated tags, is used to facilitate browsing and search process of the tags. Sinclair and Cardew-Hall [48] discuss situations when using tag clouds is better than doing search. They conducted an experiment, giving participants the option of using a tag cloud or a traditional search interface to answer various questions. They found that participants preferred the search interface if the information need is specific while they preferred the tag cloud if the information-seeking task was more general. It is partly because tags are good at broad categorization, as opposed to specific characterization of content. In total, the number of questions answered using the tag cloud was greater than those answered using the search box. The tag cloud provides a visual summary of the contents of the database. A number of participants commented that tag clouds gave them an idea of where to begin their information seeking. The tag cloud helps the information seeker to get familiar with the domain and allows more focused queries based on the information gained from browsing. It appears that scanning the tag cloud requires less cognitive load than formulating specific query terms.

Tag clouds have their disadvantages too. First, tag clouds obscure useful information by being skewed towards what is popular. Second, answering a question using the tag cloud required more queries per question than the search box. Third, many of the participants commented that the tag cloud did not allow them to narrow their search enough to answer the given questions. On average, roughly half of the articles in the dataset remain inaccessible from the tag cloud. Most tagging systems mitigate this by including tag links when viewing individual articles, thus exposing some of the less popular tags. However, this is not necessarily useful when someone is seeking specific information.

Millen et al. [34] did experiments on Dogear system where clicking a tag leads to a view of bookmarks that are associated with that tag. They found that the most frequent way to browse bookmarks is by clicking on another person's name, followed by browsing bookmarks by selecting a specific tag from the system-wide tag cloud. It is considerably less common for a user to select a tag from another user's tag cloud and very less chances of using more advanced browsing of tag intersections.

5.2 Tag Selection for Tag Clouds

Since there is only limited display space for tags in tag clouds, how to select the appropriate tags is a challenging task. Hassan-Montero and Herrero-Solana [18] describe a system for bi-dimensional visualization of tag clouds. Tag selection is based on usefulness as determined by: (1) its capacity to represent each resource as compared to other tags assigned to the same resource, (2) the volume of covered resources as compared to other tags, (3) its capacity to cover these resources less covered by other tags. Semantic relationships among tags are defined in terms of their similarity, quantified by means of the Jaccard coefficient. K-means clustering is then applied on tag similarity matrix, with an apriori chosen number of clusters and a fixed number of selected relevant tags. They apply Multidimensional Scaling, using Pearson's correlation as the similarity function, on a tag-to-tag correlation matrix. MDS creates a bi-dimensional space, which is then visualized through a fish-eye system. Alphabetical-based schemes are useful for know-item searching, i.e., when user knows previously what tag she is looking for, such as when user browses her personal tag cloud. They propose a tag cloud layout based on the assumption that clustering techniques can improve tag clouds' browsing experience. The display method is similar to traditional tag cloud layout, with the difference that tags are grouped with semantically similar tags, and likewise clusters of tags are displayed near semantically similar clusters. Similar tags are horizontally neighbors, whereas similar clusters are vertically neighbors. Clustering offers more coherent visual distribution of tags than traditional alphabetical arrangements, allowing to differentiate among main topics in tag cloud, as well as to infer semantic knowledge from the neighbors' relationships.

Begelman et al. [3] propose a clustering algorithm to find strongly related tags. The algorithm is based on counting the number of co-occurrences of any pair of tags and a cut-off point is determined when the co-occurrence count is significant enough to be used. To determine this cutoff point, they start from the tail on the right end and seek the point where the first derivative of the count has its first high peak (that is when the second derivative goes from positive to negative) and check if the peak was high enough. This results in a sparse matrix that represents tags, so that the value of each element is the similarity of the two tags. Using this definition of similarity, they design an inter-tag correlation network graph. They then cluster this graph using spectral bisection and modularity function.

5.3 Tag Hierarchy Generation

Beyond the flat structure of tags, hierarchical structure also exists in the tagging space. Caro et al. [8] present the tagFlake system, which supports semantically informed navigation within a tag cloud. The system organizes tags extracted from textual content in hierarchical organizations, suitable for navigation, visualization, classification and tracking. It extracts the most significant tag/terms from text documents and maps them onto a hierarchy in such a way that descendant terms are contextually dependent on their ancestors within the given corpus of documents. This provides tagFlake with a mechanism for enabling navigation within the tag space and for classification of the text documents based on the contextual structure captured by the generated hierarchy.

Li et al. [29] present Effective Large Scale Annotation Browser (ELSABer), to browse large-scale social annotation data. ELSABer helps the users browse a huge number of annotations in a semantic, hierarchical and efficient way. ELSABer has the following features: (1) the semantic relations between annotations are explored for browsing of similar resources; (2) the hierarchical relations between annotations are constructed for browsing in a top-down fashion; (3) the distribution of social annotations is studied for efficient browsing. The hierarchy structure is determined by a decision tree with the features including tag coverage, URL intersection rate, inverse-coverage rate, etc.

5.4 Tag Clouds Display Format

Tags clouds can be displayed in different formats. Bielenberg and Zacher [4] have proposed circular clouds, as opposed to the typical rectangular layout, where the most heavily weighted tags appear closer to the center. Font size and distance to the center represent the importance of a tag, but distance between tags does not represent their similarity.

Owen and Lemire [24] present models and algorithms to improve the display of tag clouds that consist of in-line HTML, as well as algorithms that use nested tables to achieve a more general two-dimensional layout in which tag relationships are considered. Since the font size of a displayed tag is usually used to show the relative importance or frequency of the tag, a typical tag cloud contains large and small text interspersed. A consequence is wasteful white space. To handle the space waste problem, the authors propose the classic electronic design automation (EDA) algorithm, min-cut placement, for area minimization and clustering in tag clouds. For the large clumps of white space, the solution is a hybrid of the classic Knuth-Plass algorithm for text justification, and a book-placement exercise considered by Skiena. The resulting tag clouds are visually improved and tighter.

5.5 Tag Evolution Visualization

Other than the text information, tags usually have the time dimension. To visualize the tag evolution process is an interesting topic. Dubinko et al. [14] consider the problem of visualizing the evolution of tags within Flickr. An animation provided via Flash in a web browser allows the user to observe and interact with the interesting tags as they evolve over time. The visualization is made up of two interchangeable metaphors - the 'river' and the 'waterfall'. The visualization provides a view of temporal evolution, with a large amount of surface data easily visible at each timestep. It allows the user to interact with the presentation in order to drill down into any particular result. It remains "situated" in the sense that the user is always aware of the current point of time being presented, and it provides random access into the time stream so that the user can reposition the current time as necessary. There are two novel contributions in their algorithm. The first is a solution to an interval covering problem that allows any timescale to be expressed efficiently as a combination of a small number of pre-defined timescales that have been pre-computed and saved in the "index" structure. The second contribution is an extension of work on score aggregation allowing data from the small number of pre-computed timescales to be efficiently merged to produce the optimal solution without needing to consume all the available data. The resulting visualization is available at Taglines[5]. In some cases, the user may seek data points that are particularly anomalous, while in other cases it may be data points that are highly persistent or that manifest a particular pattern. The authors focus on one particular notion of "interesting" data: the tags during a particular period of time that are most representative for that time period. That is, the tags that show a significantly increased likelihood of occurring inside the time period, compared to outside.

Russel [42] has proposed Cloudalicious[6], a tool to study the evolution of the tag cloud over time. Cloudalicious takes a request for a URL, downloads the tagging data from Delicious, and then graphs the collective users tagging activity over time. The y-axis shows the relative weights of the most popular tags for that URL. As the lines on the graph move from left to right, they show signs of stabilization. This pattern can be interpreted as the collective opinion of the users. Diagonal lines are the most interesting elements of these graphs as they suggest that the users doing the tagging have changed the words used to describe the site.

5.6 Popular Tag Cloud Demos

Some demos for visualizing tags are also available on the Web. Grafolicious[7] produces graphs illustrating when and how many times a URL has been bookmarked in Delicious. HubLog[8] gives a graph of related tags connected with the given tags. Although these demos gave a vivid picture of social annotations in different aspects, their goals are not to help users to browse annotations effectively. PhaseTwo[9] aims at creating visually pleasant tag clouds, by presenting tags in the form of seemingly random collections of circles with varying sizes: the size of the circle denotes its frequency. Delicious also provides its own tag cloud view[10]. Tag.alicio.us[11] operates as a tag filter, retrieving links from Delicious according to tag and time constraints (e.g., tags from this hour, today, or this week). Extisp.icio.us[12] displays a random scattering of a given user's tags, sized according to the number of times that the user has reused each tag, and Facetious[13] was a reworking of the Delicious database, which made use of faceted classification, grouping tags under headings such as "by place" (Iraq, USA, Australia), "by technology" (blog, wiki, website) and "by

[5] http://research.yahoo.com/taglines

[6] http://cloudalicio.us/tagcloud.php

[7] http://www.neuroticWeb.com/recursos/del.icio.us-graphs/

[8] http://hublog.hubmed.org/tags/visualisation

[9] http://phasetwo.org/post/a-better-tag-cloud.html

[10] http://del.icio.us/tag/

[11] http://planetozh.com/blog/2004/10/tagalicious-a-way-to-integrate-delicious/

[12] http://kevan.org/extispicious

[13] http://www.siderean.com/delicious/facetious.jsp

attribute" (red, cool, retro). Tag clouds have also been integrated inside maps for displaying tags having geographical information, such as pictures taken at a given location.

6. TAG RECOMMENDATIONS

The tagging system can recommend some tags to a user, and the user can select one of those tags or come up with a new one. Tag recommendation is not only useful to improve user experience, but also makes rich annotation available. There have been many studies on tag recommendation. Tags can be recommended based on their quality, co-occurrence, mutual information and object features.

6.1 Using Tag Quality

Tag quality can guide the tag recommendation process. The tag quality can be evaluated by facet coverage and popularity, and those tags of high quality are used for recommendation. Xu et al. [57] propose a set of criteria for tag quality and then propose a collaborative tag suggestion algorithm using these criteria to discover the high-quality tags. A good tag combination should include multiple facets of the tagged objects. The number of tags for identifying an object should be minimized, and the number of objects identified by the tag combination should be small. Note that personally used organizational tags are less likely to be shared by different users. Thus, they should be excluded from tag recommendations. The proposed algorithm employs a goodness measure for tags derived from collective user authorities to combat spam. The goodness measure is iteratively adjusted by a reward-penalty algorithm, which also incorporates other sources of tags, e.g., content-based auto-generated tags. The algorithm favors tags that are used by a large number of people, and minimizes the overlap of concepts among the suggested tags to allow for high coverage of multiple facets and honors the high correlation among tags.

6.2 Using Tag Co-occurrences

One important criterion used for tag recommendation is tag co-occurrence. Those tags co-occurring with the existing tags of the object are used for recommendation. Sigurbjörnsson and Zwol [46] present four strategies to recommend tags. These include two co-ocurrence based strategies: Jaccard similarity and an asymmetric measure $P(t_j \mid t_i) = \frac{|t_i \bigcap t_j|}{|t_i|}$. Tag Aggregation and promotion strategies are based on voting or weighted voting based on co-occurrence count. From the tag frequency distribution, they learned that both the head and the tail of the power law would probably not contain good tags for recommendation. Considered that user-defined tags with very low collection frequency are less reliable than tags with higher collection frequency, those tags for which the statistics are more stable were promoted. They define stability for a user using this intuition.
$stability_u = \frac{k_s}{k_s + abs(k_s - log(|u|))}$
Tags with very high frequency are likely to be too general for individual photos.
$descriptive_c = \frac{k_d}{k_d + abs(k_d - log(|c|))}$
The rank $rank(u, c)$ of a candidate tag c for a user u is $\frac{k_r}{k_r + r - 1}$ where r is rank of tag wrt co-occurrence. The promotion score can be defined as $promotion(u, c) = rank(u, c) \times stability(u) \times descriptive(c)$. Tag score is finally computed as $score(c) = \sum_{u \in U} vote(u, c) \times promotion(u, c)$. Tag frequency distribution follows a perfect power law, and the mid section of this power law contained the most interesting candidates for tag recommendation. They found that locations, artifacts and objects have a relatively high acceptance ratio (user acceptance of the recommended tag). However, people, groups and unclassified tags (tags that do not appear in WordNet) have relatively low acceptance ratio.

6.3 Using Mutual Information between Words, Documents and Tags

Mutual information is another criterion for tag recommendation. Song et al. [49] treat the tagged training documents as triplets of (words, documents, tags), and represent them as two bipartite graphs, which are partitioned into clusters by Spectral Recursive Embedding (SRE) and using Lanczos algorithm for symmetric low rank approximation for the weighted adjacency matrix for the bipartite graphs. Tags in each topical cluster are ranked by a novel ranking algorithm. A two-way Poisson Mixture Model (PMM) is proposed to model the document distribution into mixture components within each cluster and aggregate words into word clusters simultaneously. During the online recommendation stage, given a document vector, its posterior probabilities of classes are first calculated. Then based on the joint probabilities of the tags and the document, tags are recommended for this document based on their within-cluster ranking. The efficiency of the Poisson mixture model helps to make recommendations in linear-time in practice. Within a cluster, node ranking is defined by $Rank_i = exp(-\frac{1}{r(i)^2})$ for $r(i) \neq 0$ where $r(i) = np_i \times log(nr_i)$. N-Precision (np_i) of a node i is the weighted sum of its edges that connect to the nodes within the same cluster, divided by the total sum of edge weights in that cluster. N-recall (Nr_i)=edges associated with node i/edges associated with node i within the same cluster.

6.4 Using Object Features

Tag recommendation can also be performed using object features. E.g., the extracted content features from the images can be helpful in tag recommendation. In [31], Liu et al. propose a tag ranking scheme to automatically rank the tags associated with a given image according to their relevance to the image content. To estimate the tag relevance to the images, the authors first get the initial tag relevance scores based on probability density estimation, and then apply a random walk on a tag similarity graph to refine the scores. Since all the tags have been ranked according to their relevance to the image, for each uploaded image, they find the K nearest neighbors based on low-level visual features, and then the top ranked tags of the K neighboring images are collected and recommended to the user. In [55], Wu et al. model the tag recommendation as a learning task that considers multi-modality including tag co-occurrence and visual correlation. The visual correlation scores are derived from Visual language model (VLM), which is adopted to model the content of the tags in visual domain. The optimal combination of these ranking features is learned by the Rankboost algorithm.

7. APPLICATIONS OF TAGS

In this section, we would summarize different applications for which tags have been used. Social tagging can be useful in the areas including indexing, search, taxonomy generation, clustering, classification, social interest discovery, etc.

7.1 Indexing

Tags can be useful for indexing sites faster. Users bookmark sites launched by their friends or colleagues before a search engine bot can find them. Tags are also useful in deeper indexing. Many pages bookmarked are deep into sites and sometimes not easily linked to by others, found via bad or nonexistent site navigation or linked to from external pages. Carmel et al. [7] claim that by appropriately weighting the tags according to their estimated quality, search effectiveness can be significantly improved.

7.2 Search

Tags have been found useful for web search, personalized search and enterprise search. Schenkel et al. [43] develop an incremental top-k algorithm for answering queries using tags by social and semantic expansions. Hotho et al. [22] present a formal model and a new search algorithm for folksonomies, called FolkRank. Heymann et al. [21] analyze posts to Delicious and conclude that social bookmarking can provide search data not currently provided by other sources, though it may currently lack the size and distribution of tags necessary to make a significant impact. Though Noll et al. [35] observe that the amount of new information provided by metadata (tags, anchor text, search keywords) is comparatively low, Heckner et al. [19] found out using a survey that Flickr and Youtube users perceive tags as helpful for IR. Xu et al. [56] present a framework in which the rank of a web page is decided not only by the term matching between the query and the web page's content but also by the topic matching (using tags) between the user's interests and the web page's topics. Bao et al. [2] observe that the social annotations can benefit web search in two aspects. First, the annotations are usually good summaries of corresponding web pages. Second, the count of annotations indicates the popularity of web pages. Dmitriev et al. [13] show how user annotations can be used to improve the quality of intranet (enterprise) search.

7.3 Taxonomy generation

Tags have also been exploited for taxonomy generation. Heymann and Garcia-Molina [20] provide an algorithm for converting a large corpus of tags annotating objects in a tagging system into a navigable hierarchical taxonomy of tags. Schmitz et al. [44] discuss how association rule mining can be adopted to analyze and structure folksonomies, and how the results can be used for ontology learning and supporting emergent semantics. Folksonomies have the potential to add much value to public library catalogues by enabling clients to: store, maintain and organize items of interest in the catalogue [51].

7.4 Clustering and classification

Web objects can be classified and clustered more efficiently using tags. Ramage et al. [40] explore the use of tags in K-means clustering in an extended vector space model that includes tags as well as page text. Brooks and Montanez [6] analyze the effectiveness of tags for classifying blog entries by gathering the top 350 tags from Technorati and measuring the similarity of all articles that share a tag. Yin et al. [58] cast the web object classification problem as an optimization problem on a graph of objects and tags.

7.5 Social interesting discovery

Tags can be useful for social interest discovery. Li et al. [30] propose that human users tend to use descriptive tags to annotate the contents that they are interested in. As described in section 4, Rattenbury et al. [41] focus on the problem of extracting place and event semantics for Flickr tags. Looking at the latest place and event tags can help us discover recent popular events.

7.6 Enhanced Browsing

Zubiaga et al. [60] suggest alternative navigation ways using social tags: pivot browsing (moving through an information space by choosing a reference point to browse), popularity driven navigation (sometimes a user would like to get documents that are popular for a known tag), and filtering (social tagging allows to separate the stuff you do not want from the stuff you do want). Millen et al. [34] perform experiments to study user browsing habits in presence of tag clouds.

7.7 Integrated folksonomies

Tags from multiple folksonomies can be combined using tag co-occurrence analysis and clustering [50]. An integration of such folksonomies can help in solving the problem of sparsity of tags associated with an object. Cross-linking distributed user tag clouds can help in creating richer user profiles [52]. TAGMAS (TAG Management System) [23] is a federation system that supplies a uniform view of tagged resources distributed across a range of Web2.0 platforms. Such a system can be useful for automatic tag creation (which permit to create desktop-specific tags), folksonomy loading (which permit to import a folksonomy from a folkserver), resource annotation (where a resource can be annotated along loaded folksonomies) and resource searching (where tag-based filtering is used to locate resources regardless of where the resource is held).

8. TAGGING PROBLEMS

Though tags are useful, exploiting them for different applications is not easy. Tags suffer from problems like spamming, canonicalization and ambiguity issues. Other problems such as sparsity, no consensus, etc. are also critical. In this section, we discuss these problems and suggest solutions described in the literature.

8.1 Spamming

Spammers can mis-tag resources to promote their own interests. Wetzker et al. [54] have observed such phenomena where a single user labeled a large number of bookmarks with the same tags all referring to the same blog site. They have also observed a phenomenon where users upload thousands of bookmarks within minutes and rarely actively contribute again. They characterize the spammers as possessing these properties: very high activity, tagging objects belonging to a few domains, high tagging rate per resource, and bulk posts. To detect such spamming, they propose a new concept called diffusion of attention which helps to reduce the influence of spam on the distribution of tags without the actual need of filtering. They define the attention a tag achieves in a certain period of time as the number of users using the tag in this period. The diffusion for a tag is then given as the number of users that assign this tag for the first time. This measures the importance of an item by its capability to attract new users while putting all users on an equal footage. Every user's influence is therefore limited and a trend can only be created by user groups.

Koutrika et al. [26] study the problem of spamming extensively. How many malicious users can a tagging system tolerate before results significantly degrade? What types of tagging systems are more vulnerable to malicious attacks? What would be the effort and the impact of employing a trusted moderator to find bad postings? Can a system automatically protect itself from spam, for instance, by exploiting user tag patterns? In a quest for answers to these questions, they introduce a framework for modeling tagging systems and user tagging behavior. The framework combines legitimate and malicious tags. This model can study a range of user tagging behaviors, including the level of moderation and the extent of spam tags, and compare different query answering and spam protection schemes. They describe a variety of query schemes and moderator strategies to counter tag spam. Particularly, they introduce a social relevance ranking method for tag search results that takes into account how often a user's postings coincide with others' postings in order to determine their "reliability". They define a metric for quantifying the "spam impact" on results. They compare the various schemes under different models for malicious user behavior. They try to understand the weaknesses of existing systems and the magnitude of the tag spam problem. They also make predictions about which schemes will be more useful and which malicious behaviors will be more disruptive in practice.

8.2 Canonicalization and Ambiguities

Ambiguity arises in folksonomies because different users apply terms to documents in different ways. Acronyms can also lead to ambiguities. Users often combine multiple words as a single tag, without spaces, e.g., 'vertigovideostillsbbc' on Flickr. Currently, tags are generally defined as single words or compound words, which means that information can be lost during the tagging process. Single-word tags lose the information that would generally be encoded in the word order of a phrase. There is no synonym or homonym control in folksonomies. Different word forms, plural and singular, are also often both present. Folksonomies provide no formal guidelines for the choice and form of tags, such as the use of compound headings, punctuation, word order. In addition, the different expertise and purposes of tagging participants may result in tags that use various levels of abstraction to describe a resource.

Guy and Tonkin [16] point out the existence of useless tags due to misspellings, bad encoding like an unlikely compound word grouping (e.g., TimBernersLee); tags that do not follow convention in issues such as case and number; personal tags that are without meaning to the wider community (e.g., mydog); single use tags that appear only once in the database(e.g., billybobsdog), symbols used in tags. Conventions have become popular, such as dates represented according to the ISO standard (eg. 20051201 for "1st December, 2005") and the use of the year as a tag. One wildly popular convention is geotagging, a simple method of encoding latitude and longitude within a single tag; this represented over 2% of the total tags sampled on Flickr. A common source of "misspelt" tags was in the transcoding of other alphabets or characters.

Zubiaga [60] suggests a solution to the canonicalization problem. To merge all forms of the same tag, the system can rely on a method like that by Librarything. This site allows users to define relations between tags, indicating that some of them have the same meaning. In his blog entry, Lars Pind [37] has suggested various ways to solve canonicalization problem, including the following: (1) suggest tags for user, (2) find synonyms automatically, (3) help user use the same tags that others use, (4) infer hierarchy from the tags, and (5) make it easy to adjust tags on old content. Quintarelli [39] mentions that the system can have a correlation feature that, given a tag, shows related tags, i.e., tags that people have used in conjunction with the given tag to describe the same item. Guy and Tonkin [16] suggest educating the users, simple errorchecking in systems when tags are entered by users, making tag suggestions (synonyms, expansion of acronyms etc.) when users submit resources (e.g., using Scrumptious, a recent Firefox extension, offers popular tags from Delicious for every URL). They also suggest creation of discussion tools through which users can share reasons for tagging things in a certain way. More understanding of who is submitting certain tags could possibly alter personal rating of posts by other users.

8.3 Other Problems

There are many other problems related to social tagging, including sparsity, no consensus and search inefficiency. Sparsity is related to the annotation coverage of the data set. Bao et al. [2] point out that certain pages may not be tagged at all. Users do not generally associate tags to newly emerging web pages or web pages that can be accessed easily from hub pages, or uninteresting web pages. Halpin et al. [17] point that users may not reach a consensus over the appropriate set of tags for a resource leading to an unstable system. As Golder and Huberman [15] suggest, changes in the stability of such patterns might indicate that groups of users are migrating away from a particular consensus on how to characterize a site and its content or negotiating the changing meaning of that site. Quintarelli [39] points out that tags have no hierarchy. Folksonomies are a flat space of keywords. Folksonomies have a very low findability quotient. They are great for serendipity and browsing but not aimed at a targeted approach or search. Tags do not scale well if you are looking for specific targeted items.

9. CONCLUSION AND FUTURE DIRECTIONS

In this work, we surveyed social tagging with respect to different aspects. Tags are taking on a new meaning as other forms of media like microblogs are gaining popularity. Below we mention a few aspects which can be a part of future research.

9.1 Analysis

Most current research on tagging analysis focus on one single tag stream itself. However, as the type of user generated content evolves, tags may be different and related to different kinds of user generated data, such as microblogs and query logs. For example, How does tag growth differ in microblogs versus that for bookmarks and images. Tagging models for microblogs can be quite different from other tagging models. E.g., certain tags reach a peak on twitter quite unexpectedly. These tags don't relate to any specific events. Such varying degree of social influence when a pseudo event happens hasn't been captured by any of the tagging models, yet.

9.2 Improving system design

Current tagging systems only support a type of tags and researchers have developed mechanisms to extract hierarchical structures (ontology) from this flat tagging space. Systems can provide more functionality like hierarchical tags, say (programming/java), multi-word tags. A tagging system can also support a tag discussion forum where users can debate about the appropriateness of a tag for a resource. Structured tags can also be supported, i.e., allow people to tag different portions of a web page with different tags and assign key=value pairs rather than just "values". E.g., person="Mahatma Gandhi", location="Porbandar", year="1869", event="birth". By adding more such functionality into the system, we can expect that a more meaningful semantic structure could be extracted.

9.3 Personalized tag recommendations

Is the user a describer or organizer? What is the context? Is she tagging on sets in Flickr or just photos in the photostream (i.e., context within the tagging site itself)? Based on her history, what is the probability that she would choose a new tag? What are the words used in her previous tags, words used in her social friends' tags? Given some tags tied to a resource, we can identify whether users prefer to repeat tags for this resource or do they like to put on new tags. Using this we can vary the tag history window size shown with the resource. Apart from tag recommendation, a recommendation system can also recommend related resources once a user selects a particular tag.

9.4 More applications

There are also interesting applications which are worth exploring. E.g., (1) Tagging support for desktop systems using online tags. (2) Geographical/demographical analysis of users' sentiments based on the tags they apply to products launched at a particular location. (3) Mashups by integrating resources with same/similar tags. (4) Establishing website trustworthiness based on what percent of the keywords mentioned in the <meta> tag are actual tags for web page bookmarks. (5) Summary generation using tags with NLP. (6) Intent detection and behavioral targeting based on user history of tags.

9.5 Dealing with problems

Sparsity, canonicalization, ambiguities in tags still remain as open problems. More work needs to be done to come up with solutions to effectively deal with them. Also, certain tags get outdated. E.g., a camera model may be marked as 'best camera'. But after two years, it no longer remains the 'best camera'. How can we clean up such kind of tag rot? Similarly, tags that haven't been repeated by another user within a time window, can be considered as personal tags and can be removed from public view.

10. REFERENCES

[1] Morgan Ames and Mor Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Conference on Human Factors in Computing Systems, CHI 2007*, Sam Jose, CA, April 2007.

[2] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA, 2007. ACM.

[3] Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space, 2006.

[4] K. Bielenberg. Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation. Master's thesis, 2005.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] Christopher H. Brooks and Nancy Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM Press.

[7] David Carmel, Haggai Roitman, and Elad Yom-Tov. Who tags the tags?: a framework for bookmark weighting. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1577–1580, New York, NY, USA, 2009. ACM.

[8] Luigi Di Caro, K. Seluk Candan, and Maria Luisa Sapino. Using tagflake for condensing navigable tag hierarchies from tag clouds. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *KDD*, pages 1069–1072. ACM, 2008.

[9] Ciro Cattuto, Andrea Baldassarri, Vito D. P. Servedio, and Vittorio Loreto. Vocabulary growth in collaborative tagging systems, Apr 2007.

[10] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences (PNAS)*, 104(5):1461–1464, January 2007.

[11] Klaas Dellschaft and Steffen Staab. An epistemic dynamic model for tagging systems. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 71–80, New York, NY, USA, 2008. ACM.

[12] Nicholas Diakopoulos and Patrick Chiu. Photoplay: A collocated collaborative photo tagging game on a horizontal display. preprint (2007) available at http://www.fxpal.com/publications/FXPAL-PR-07-414.pdf.

[13] Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita. Using annotations in enterprise search. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 811–817, New York, NY, USA, 2006. ACM.

[14] Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM Press.

[15] Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems, Aug 2005.

[16] Marieke Guy and Emma Tonkin. Folksonomies: Tidying up tags? *D-Lib Magazine*, 12, Jan 2006.

[17] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2007. ACM.

[18] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *InScit2006: International Conference on Multidisciplinary Information Sciences and Technologies*, 2006.

[19] M. Heckner, M. Heilemann, and C. Wolff. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, San Jose, CA, USA, May 2009.

[20] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University, April 2006.

[21] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA, 2008. ACM.

[22] Andreas Hotho, Robert Jschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. *The Semantic Web: Research and Applications*, pages 411–426, 2006.

[23] Jon Iturrioz, Oscar Diaz, and Cristobal Arellano. Towards federated web2.0 sites: The tagmas approach. In *Tagging and Metadata for Social Information Organization Workshop, WWW07*, 2007.

[24] Owen Kaser and Daniel Lemire. Tag-cloud drawing: Algorithms for cloud visualization, May 2007.

[25] Margaret E. I. Kipp and Grant D. Campbell. Patterns and inconsistencies in collaborative tagging systems : An examination of tagging practices. In *Annual General Meeting of the American Society for Information Science and Technology.* American Society for Information Science and Technology, November 2006.

[26] Georgia Koutrika, Frans A. Effendi, Zolt'n Gyöngyi, Paul Heymann, and Hector G. Molina. Combating spam in tagging systems: An evaluation. *ACM Trans. Web*, 2(4):1–34, 2008.

[27] Christian Krner. Understanding the motivation behind tagging. ACM Student Research Competition - Hypertext 2009, July 2009.

[28] Liz Lawley. social consequences of social tagging. Web article, 2005.

[29] Rui Li, Shenghua Bao, Yong Yu, Ben Fei, and Zhong Su. Towards effective browsing of large scale social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 943–952, New York, NY, USA, 2007. ACM.

[30] Xin Li, Lei Guo, and Yihong E. Zhao. Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 675–684, New York, NY, USA, 2008. ACM.

[31] Dong Liu, Xian S. Hua, Linjun Yang, Meng Wang, and Hong J. Zhang. Tag ranking. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 351–360, New York, NY, USA, April 2009. ACM.

[32] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, toread. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[33] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication*, December 2004.

[34] David R. Millen and Jonathan Feinberg. Using social tagging to improve social navigation. In *Workshop on the Social Navigation and Community based Adaptation Technologies*, 2006.

[35] Michael G. Noll and Christoph Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. *WI/IAT '08: Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:640–647, 2008.

[36] Simon Overell, Börkur Sigurbjörnsson, and Roelof van Zwol. Classifying tags using open content resources. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 64–73, New York, NY, USA, 2009. ACM.

[37] Lars Pind. Folksonomies: How we can improve the tags. Web article, 2005.

[38] Peter Pirolli. Rational analyses of information foraging on the web. *Cognitive Science*, 29(3):343–373, 2005.

[39] Emanuele Quintarelli. Folksonomies: power to the people.

[40] Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector G. Molina. Clustering the tagged web. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63, New York, NY, USA, 2009. ACM.

[41] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIRIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, New York, NY, USA, 2007. ACM Press.

[42] Terrell Russell. cloudalicious: folksonomy over time. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 364–364, New York, NY, USA, 2006. ACM.

[43] Ralf Schenkel, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane X. Parreira, and Gerhard Weikum. Efficient top-k querying over social-tagging networks. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 523–530, New York, NY, USA, 2008. ACM.

[44] Christoph Schmitz, Andreas Hotho, Robert Jschke, and Gerd Stumme. Mining association rules in folksonomies. In *Data Science and Classification*, pages 261– 270. Springer, 2006.

[45] Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. tagging, communities, vocabulary, evolution. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 181–190, New York, NY, USA, November 2006. ACM.

[46] Brkur Sigurbjrnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th International Conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.

[47] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[48] James Sinclair and Michael Cardew-Hall. The folksonomy tag cloud: when is it useful? *J. Inf. Sci.*, 34(1):15–29, 2008.

[49] Yang Song, Ziming Zhuang, Huajing Li, Qiankun Zhao, Jia Li, Wang C. Lee, and C. Lee Giles. Real-time automatic tag recommendation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 515–522, New York, NY, USA, 2008. ACM.

[50] Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *ESWC '07: Proceedings of the 4th European conference on The Semantic Web*, pages 624–639, Berlin, Heidelberg, 2007. Springer-Verlag.

[51] Louise F. Spiteri. Structure and form of folksonomy tags: The road to the public library catalogue. *Webology*, 4(2, Artikel 41), 2007.

[52] Martin Szomszor, Ivan Cantador, and Harith Alani. Correlating user profiles from multiple folksonomies. In *ACM Confrence on Hypertext and Hypermedia*, June 2008.

[53] Csaba Veres. The language of folksonomies: What tags reveal about user classification. In *Natural Language Processing and Information Systems*, volume 3999/2006 of *Lecture Notes in Computer Science*, pages 58–69, Berlin / Heidelberg, July 2006. Springer.

[54] Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Proceedings of the ECAI 2008 Mining Social Data Workshop*, pages 26–30. IOS Press, 2008.

[55] Lei Wu, Linjun Yang, Nenghai Yu, and Xian S. Hua. Learning to tag. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 361–370, New York, NY, USA, 2009. ACM.

[56] Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, and Yong Yu. Exploring folksonomy for personalized search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 155–162, New York, NY, USA, 2008. ACM.

[57] Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *WWW2006: Proceedings of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, 2006.

[58] Zhijun Yin, Rui Li, Qiaozhu Mei, and Jiawei Han. Exploring social tagging graph for web object classification. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 957–966, New York, NY, USA, 2009. ACM.

[59] Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and C. Lee Giles. Exploring social annotations for information retrieval. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 715–724, New York, NY, USA, 2008. ACM.

[60] Arkaitz Zubiaga. Enhancing navigation on wikipedia with social tags. In *Wikimania '09*, 2009.